

Fig. S1: MDS visualizations of OTU clusters for species in the *Claroideoglomus/Entrophospora* group; 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). Calculating the adjusted Rand index is not applicable to this group because due to the incomplete lineage sorting of sequence variants it displays.

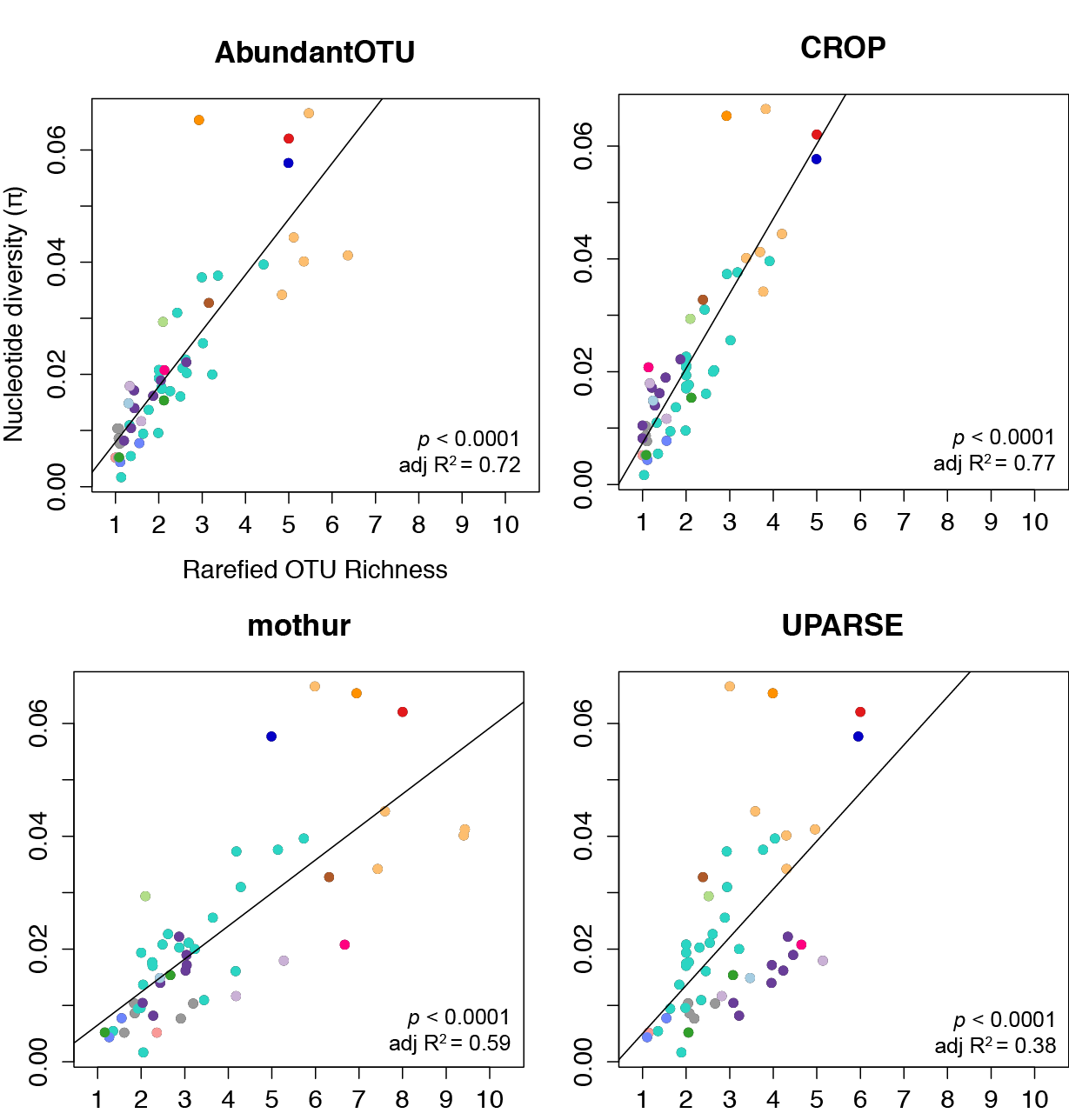


Fig. S2: Linear regressions between rarefied OTU richness and nucleotide diversity for each of the four clustering methods used. Each point represents values for sequences from an individual sequencing barcode and are colored by genus following Fig. 1.

Supplemental Methods

*Dataset of rRNA sequences from AM fungal species*

Our dataset comprised sequences from three sources and was a combination of ‘reference’ and ‘test’ sequences. The reference sequences were themselves obtained from two sources: 1) a previously published and hand-curated alignment (1), retaining only the sequences that contained the portion of the phylogenetically informative D2 region of the nuclear large subunit (LSU; 28S) rRNA gene that corresponded with the test sequences, and 2) supplemental sequences with confident species attributions obtained from GenBank to expand the phylogenetic breadth of the reference dataset. The supplemental sequences were added to the existing alignment using the ‘add’ function in MAFFT (2).

The test sequences were obtained from resting spores of 21 morphologically-defined species of AM fungi that were isolated from soil-based cultures. Geographic isolates of four species were used to determine sub-species levels of sequence variation and replicate DNA extractions of eight species were used to determine variability between replicate biological samples. Details of spore isolation and DNA extraction procedures are described in more detail in Kaonongbua, Morton and Bever (3). Briefly, spores were cleaned by sonication and either single or multiple spores were crushed in TE buffer and then heated to 100°C before freezing for use as a template for all PCR reactions. Forward PCR primer LROR (4) was used with reverse primer FLR2 (5) to amplify a roughly 850bp fragment of the LSU rRNA gene. PCR products were then pyrosequenced (454, Roche) using primer FLR2 to give sequences of the D2 region of the LSU rRNA gene.

The raw test sequences (188,713 sequences) were subjected to stringent quality controls. First, sequences with any undetermined bases (Ns), with insertions or deletions relative to the PCR primer, or more than one substitution relative to the PCR primer were removed before trimming the primer from the remaining sequences. Sequences were truncated to 350bp and then quality filtered using the UPARSE (6), allowing a maximum of 1 expected error per sequence. All remaining reads were de-replicated to remove identical sequences, and chimeric sequences were removed using UCHIME denovo (7). After these quality control procedures, 9,094 unique test sequences remained. These were then added to the de-replicated reference sequences for use in the multidimensional scaling (MDS) as well as the CROP and UPARSE clustering datasets (10,081 sequences); the unique test sequences were then re-replicated (preserving the original abundance of each unique non-chimeric test sequence) and added to the reference sequences to form the dataset for use in the AbundantOTU and mothur clustering datasets (51,543 sequences). The reference sequences were included to represent AM fungal clades not sampled with the test sequences, and to validate test sequences for a given species using sequences obtained from another source. However, only the test sequences were used for all analyses of sequence variation and OTU clustering accuracy.

*Visualization of sequence variation*

Sequence differences were visualized using a novel MDS visualization algorithm within the DACIDR sequence clustering pipeline (8) that represents each sequence by a point in three-dimensional space. To compare the resulting visualizations of sequence variation with their evolutionary relationships, we interpolated a phylogenetic tree into the MDS visualization using a neighbor-joining algorithm we developed previously (9). For clarity, not all sequences were included in the phylogeny; instead the dataset was pre-clustered using AbundantOTU with a 99% sequence similarity cut-off. The resulting sequences still represented all genera present in the initial dataset. We aligned the sequences with MUSCLE (10), and created an unrooted maximum likelihood phylogenetic tree using RAxML (11).

To better evaluate the sequence similarity between species from specific clades with different amounts of sequence variation, we constructed a rooted phylogenetic tree and a heatmap of pairwise sequence divergence for each clade. The phylogenetic tree was constructed with RAxML (11) using all aligned sequences from the clade and one sequence from the most closely related clade in the dataset (from Figure 1) as an outgroup. On the resulting phylogenetic tree, branches containing sequences from the same species were collapsed for clarity and were colored according to species. Heatmaps of pairwise sequence divergence were made by taking a random sub-sample of sequences from each species, aligning them using MUSCLE (10) and then calculating the pairwise genetic distance following the Tamura-Nei model of nucleotide substitution (12) using MEGA6 (13).

*Evaluation of phylogenetic signal in patterns of sequence variation*

For each 454 barcode, which represents sequences from different species, geographic isolates, or replicate DNA extractions, we estimated the per-site nucleotide diversity (π) by first aligning all sequences for that barcode using MUSCLE (39). From each of these alignments by barcode we then took 10 random subsamples of 85 sequences each (the smallest number of sequences in the dataset for any species: *Diversispora spurca*). For each of the 10 subsamples per barcode, π was calculated using:

(1)

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, *ni* is the number of nucleotides present in the alignment at each position with a SNP, and *FAi*, *FTi*, *FCi*, and *FGi* are the frequencies of nucleotides A, T, C, and G, respectively in the alignment at each position with a SNP. We then averaged the π estimates from the 10 subsamples for each barcode and used those values of π for all analyses.

We clustered the dataset into OTUs using AbundantOTU, mothur, and UPARSE with a 97% sequence similarity threshold using default settings; for CROP, we used similarity levels meant to approximate 97%. For each clustering method, we calculated the number of OTUs in which sequences from each species appeared (OTU richness), and the Shannon diversity index of the number of sequences from each species contained in each OTU (OTU diversity). We corrected OTU richness for sequence number by rarefying the OTU richness values of all samples to the minimum number of sequences per species in the dataset (85 sequences) using the rarefy function in the ‘vegan’ package in R.

To determine the extent to which nucleotide diversity, rarefied OTU richness, and OTU diversity vary across different taxonomic groups, we tested the magnitude and significance of variance components corresponding to AM fungal genera, species, and geographic isolates with mixed models using the Proc Mixed routine in SAS software (SAS Institute Inc. Cary, NC). We followed the consensus AM fungal genus names proposed by Redecker*, et al.* (43), except for *Claroideoglomus* and *Entrophospora*, which we grouped together (see results and discussion). Total variance was calculated as the sum of the variances explained by differences among geographic isolates, species, and genera; if at least one of those variance components was significantly different from zero, then we also considered the total variance to be significantly different. For rarefied OTU richness and OTU diversity we repeated this analysis for each of the four clustering methods.

To test for phylogenetic signal in the variation of nucleotide diversity, rarefied OTU richness, and OTU diversity we adapted an approach developed to estimate the heritability using genetic markers (14). We estimated phylogenetic heritability by regressing pairwise phylogenetic distance against the pairwise cross product of sequence diversity, defined as:

(2)

Where *yi*and *yj*are the trait (sequence diversity, rarefied OTU richness, or OTU diversity) values for each of the samples (sequences from each barcode) in the pairwise comparison, is the mean trait value for all the samples, and *V* is the unbiased variance of the trait value for all of the samples. The phylogenetic heritability is estimated using the regression:

(3)

Where *ri,j* is the pairwise phylogenetic distance for each of the samples in the pairwise comparison, *h2* is the phylogenetic heritability, *rei,j*is the pairwise correlation of each of the samples due to environmental factors, and *ei,j* is measurement error. Because *rei,j*and *ei,j* are assumed to be independent of *ri,j,* the phylogenetic heritability can be calculated as:

(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. The pairwise phylogenetic distances were calculated using the ‘cophenetic’ function in the ‘ape’ library of R on a rooted phylogeny made with RAxML from an alignment of extended sequences (~675 bp) for each of the species in the test dataset that were collected either from (1) (19 species), from GenBank: *Cetraspora pellucida* and *Rhodotorula hordea* (outgroup), or sequenced by us (*E. infrequens*). Geographic isolates of the same species were added to the phylogenetic distance matrix as entries with values of zero.

*Assessing OTU accuracy compared to 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. In this way, we were able to visually compare the distribution of OTUs within different genera and species, and to determine how OTUs produced by the four clustering methods differed in their sequence membership. We evaluated the accuracy of OTUs produced by each clustering method compared to the known species composition by calculating the adjusted Rand index (45, 46) using the phyclust package in R with the null model assumption that sequences from each species would be assigned to a single OTU.

Supplemental Table 1 – See Excel spreadsheet attached to email

References

1. 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.

2. Katoh K & Frith MC (2012) Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics* 28(23):3144-3146.

3. 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.

4. Bunyard BA, Nicholson MS, & Royse DJ (1994) A Systematic Assessment of Morchella Using RFLP Analysis of the 28S ribosomal RNA gene. *Mycologia* 86(6):762-772.

5. Trouvelot S, van Tuinen D, Hijri M, & Gianinazzi-Pearson V (1999) Visualization of ribosomal DNA loci in spore interphasic nuclei of glomalean fungi by fluorescence in situ hybridization. *Mycorrhiza* 8(4):203-206.

6. Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Meth* 10(10):996-998.

7. Edgar RC, Haas BJ, Clemente JC, Quince C, & Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27(16):2194-2200.

8. Ruan Y*, et al.* (2012) DACIDR: deterministic annealed clustering with interpolative dimension reduction using a large collection of 16S rRNA sequences. in *ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pp 329-336.

9. 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.

10. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32(5):1792-1797.

11. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*.

12. Tamura K & Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular biology and evolution* 10(3):512-526.

13. Tamura K, Stecher G, Peterson D, Filipski A, & Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution* 30(12):2725-2729.

14. Ritland K (2000) Marker-inferred relatedness as a tool for detecting heritability in nature. *Molecular Ecology* 9(9):1195-1204.