Supplement S3 - Detailed DNA metabarcoding methods

Baker CCM, Martins DJ, Pelaez JN, Billen JPJ, Pringle A, Frederickson ME & NE Pierce (2017). Distinctive fungal communities in an obligate African ant plant mutualism. *Proceedings of the Royal Society B* 20162501. http://dx.doi.org/10.1098/rspb.2016.2501

DNA extractions – domatium and leaf samples

We prepared domatium and leaf samples for DNA extraction by placing each sample in a sterile 1.5mL microcentrifuge tube, removing excess preservative with a pipette, and rinsing the sample with molecular grade water. For fresh leaf samples and the old leaflet samples from CN domatia, we counted out 40 leaves for the extraction. For the fibrous particle samples from TP domatia and the carton lamellae from CM domatia, we measured out 80 to 100mg wet weight. We used the whole sample wherever less material was available.

We homogenized each sample in an MPBio Lysing Matrix A tube (including ceramic sphere; MP Biomedicals LLC, Santa Ana CA) using an MPBio FastPrep-24 benchtop homogenizer. Each tube contained the sample plus 1000μL of Qiagen Buffer AP1 with 4μL Qiagen Proteinase K (20mg/mL; Qiagen Inc, Valencia CA) added. Other researchers have found Proteinase K to be important for successful extraction from samples preserved in ammonium sulfate [3]. We homogenized samples for 40 sec at speed setting 6.0, removed the samples to ice for 5 min, then homogenized for a further 40 sec at the same speed.

We incubated the homogenized samples for 10 min at $60-65^{\circ}$C, inverting the tubes 2-3 times during incubation. We further incubated the tubes overnight at 55°C in a shaking incubator.

We continued the extraction protocol the next day, following the Qiagen DNeasy Plant Mini Kit protocol with the following modifications. We added 325μL Buffer AP2 to the lysate, since the lysate volume was larger than that in the original protocol. After precipitation, we split the lysate for each sample into two portions and passed each portion through a separate QIAshredder column. After adding Buffer AP3/E, we passed the combined flow-through from both QIAshredder columns through a single DNEasy Mini spin column. We washed the adsorbed DNA with a single 500μL volume of Buffer AW, and eluted with two 100μL volumes of Buffer AE.

DNA extracts were quantified with a NanoDrop and a Qubit fluorometer, precipited in ethanol and resuspended in Buffer AE at 20ng/μL, or at a lower concentration if required in order to have a minimum 20μL once resuspended.

DNA extractions – alate samples

Prior to DNA extraction, we surface sterilized alates to minimize any contribution from fungi present on the outside, by rinsing for 1 min in 100% ethanol, followed by 1 min in 10% bleach, and a final 1 min in 100% ethanol. Flame sterilized tools were used to handle the alates at all times.

We extracted DNA from all 32 alates following a standard phenol-chloroform extraction protocol. Using forceps and a scalpel blade, alate heads were removed and placed in individual 2mL Sarstedt tubes with five 0.5mm glass beads. We homogenized these samples in a Mini-Beadbeater-8 (BioSpec Products, Bartlesville OK) at full speed for 1 min. We then added 400μL CTab buffer, and incubated samples overnight at $60-65^{\circ}$C.

Following incubation, we added 400μL of 25:24:1 phenol:chloroform:isoamyl alcohol and repeatedly inverted the tubes for 1 min to mix. We then centrifuged samples for 15 min at 13000 rpm. We removed

300μL supernatant to a new tube and added 300μL of 24:1 chloroform:isoamyl alcohol before again mixing for 1 min and centrifuging for 15 min at 13000 rpm. We transferred 200μL supernatant to a new tube, and added 500μL of 100% ethanol, 75μL sodium acetate, and 3μL glycogen to precipitate DNA.

After precipitating overnight at 20°C, we centrifuged samples for 15 min at 13000 rpm and removed supernatant. We then washed the pellet in 500μL of 70% ethanol and centrifuged the samples again for 10 min at 13000 rpm. We removed the ethanol with a pipette, air dried the pellets and resuspended DNA in 100μL of molecular grade water. Extracts were quantified with a NanoDrop and a Qubit fluorometer. Extracts were stored at −20°C until PCR amplification and sequencing.

PCR amplifications – alate samples

Since the alate samples were small and likely varied in the amount of fungal material that they contained – particularly if some were not carrying an infrabuccal pellet – we used PCR amplifications to assess the presence of fungal material in each alate sample. PCR products were visualized using agarose gel electrophoresis and scored for either successful or unsuccessful amplification. Each reaction consisted of 2.5μL Omega BioTek 10X buffer, 1.0μL $MgCl_2$ at 25mM, 0.25μL dNTPs at 25mM, 1.2μL each primer at 10μM, 1U Omega BioTek Taq polymerase, 1μL DNA template and molecular grade H_2O to 25μL. Reaction conditions were 2 min at 94°C, followed by 34 cycles of 35 sec at 95°C / 55 sec at 55°C / 45 sec at 72°C, and a final 7 min at 72°C. We used the primers ITS1F (`5'-CTTGGTCATTTAGAGGAAGTAA-3'`; see [9]) and ITS4 (`5'-TCCTCCGCTTATTGATATGC-3'`; see [24]) to capture the broad fungal community while excluding non-fungal taxa.

Multiplexed amplicon pyrosequencing – all samples

Amplification and pyrosequencing were performed by Research and Testing Laboratory (RTL), Lubbock, TX. We sent 75 domatium and leaf samples for sequencing, plus the 16 CN alates and 5 TP alates that we successfully PCR amplified and that passed NanoDrop/Qubit quality checks.

Samples were first amplified using forward and reverse fusion primers. The forward primer was made up of the Roche A linker (454 Life Sciences, Branford CT), an 8bp multiplex identifier and the ITS1F fungal primer (`5'-CTTGGTCATTTAGAGGAAGTAA-3'`; see [9]). The reverse primer consisted of a biotin molecule, the Roche B linker, and the ITS4 primer (`5'-TCCTCCGCTTATTGATATGC-3'`; see [24]). PCRs comprised 1μL of each primer at 5μM, 1μL of extract, Qiagen HotStarTaq Master Mix plus water to a total volume of 25μL. RTL performed reactions on ABI Veriti Thermo Cyclers (Applied Biosytems, Carlsbad CA) with the following conditions: 5 min at 95°C, followed by 35 cycles of 30 sec at 94°C / 40 sec at 54°C / 60 sec at 72°C, and a final 10 min at 72°C. PCR products were then pooled equimolar.

Pooled PCR products were cleaned with Diffinity RapidTips (Diffinity Genomics, West Henrietta NY), and size selected using Agencourt AMPure XP beads (BeckmanCoulter, Indianapolis IN) following Roche 454 protocols. RTL then hybridized 150ng of DNA to Dynabeads M-270 (Life Technologies, Grand Island NY) to create single stranded DNA, again following Roche 454 protocols. Single stranded DNA was diluted and used in emulsion PCR reactions, which were subsequently enriched. Sequencing was performed on a Roche GS FLX 454 pyrosequencer with Titanium chemistry following standard manufacturer protocols.

Bioinformatics – all samples

RTL provided demultiplexed sff files generated using the Roche sffinfo tool. The demultiplexing operation filters out sequences whose terminal regions (i.e. primer, linker or barcode sequences) fail to match the known set of valid sequences, and thus functioned as a basic quality filter.

We processed the prefiltered sequence data ourselves using a combination of software tools. We denoised sequences with Ampliconnoise [19] as implemented in the QIIME 1.6 bioinformatics pipeline [6]. We then isolated the ITS1 region from our reads using an open-source software utility provided by Nilsson *et al.*, as OTU clustering may be distorted by the inclusion of conserved flanking regions [16]. Because our unidirectional sequencing started from the end of 18S, virtually all sequences contained ITS1; any remaining sequence (i.e. 18S, 5.8S or ITS2) was discarded from our analysis. Previous studies suggest that ITS1 and ITS2 yield similar results for this kind of analysis [4], so we do not consider biases arising from our choice of sequencing direction likely to be highly misleading.

We picked OTUs based on the ITS1 fragments using uclust [8] in QIIME with a similarity threshold of 95%. We used the full dataset for the clustering – i.e. the 75 domatium and leaf samples, plus the 12 CN alate and 3 TP alate samples that yielded more than a minimum 300 sequences.

For comparison, we also tried several variants on this workflow, including open reference OTU picking in QIIME and ESPRIT complete-linkage based hierarchical clustering [23] instead of uclust, different similarity thresholds, and clustering with full-length sequences rather than just ITS1. Results were broadly similar to those presented here.

We assigned putative taxonomic descriptions to a representative sequence from each OTU using blast via QIIME's assign_taxonomy.py script. In short, we downloaded nucleotide (db=nuccore) sequences from NCBI with the query

```
"fungi[Organism] NOT (environmental sample[filter] OR metagenomes[orgn])"
```

to help exclude unidentified environmental sequences, and built a blast database from those sequences. We then used our own Python script to parse the NCBI taxonomy database and return a QIIME-compatible taxonomy mapping file for our blast database. Our Python code is freely available and may be applied to any fasta file using NCBI GI numbers as identifiers (i.e. including any sequence set downloaded from the NCBI using gquery or e-direct).

Data analysis – domatium and leaf samples

We performed statistical analyses of the domatium and leaf OTU data using a combination of QIIME, the statistics package R [20] and the R packages vegan [17], biom [14], RcolorBrewer [15], spatstat [2] , randomForest [11], plyr [25], bipartite [7] and shape [21], plus our own custom scripts.

We used a rarefaction curve approach to assess differences in taxon richness and diversity between groups of samples. We resampled our domatium and leaf sample OTU table 1000 times at predefined per-sample depths. For each sample and for each sampling depth, we calculated the mean number of observed species across the 1000 resamplings. We then used these averages to find group means and standard errors at each sampling depth, so that our standard errors primarily reflect between-sample variation.

To formally test for differences in diversity evident in the rarefaction curves, we used simple parametric statistical tests after rarefying each sample to 1000 sequences, since sampling depth may affect observed taxonomic richness and beta diversity comparisons [10]. Although some authors argue that rarefying is statistically inefficient [13], we were relatively unconcerned about type II errors in this dataset, as effect sizes appeared reasonably large. Rarefying to 1000 sequences allowed us to retain all domatium and leaf samples in the analysis, since the minimum of sequences per sample was 1119 sequences. As the rarefaction curves do not cross, our choice of rarefaction depth is unlikely to have much influence over our qualitative assessment of alpha diversity, which is our main interest here.

We used distance-based ordination to assess variation in fungal community composition within and among our sample types and sampling locations. Although phylogenetic distances such as UniFrac are commonly used in work on bacterial communities and may be more informative [12], the ITS region is not easily aligned among highly diverged taxa such as those in our samples [18]. Moreover, no multiple alignment is

available as a basis for alignment akin to the Silva or Greengenes datasets for bacterial 16S. We therefore chose to use Sørensen [22] and Bray-Curtis [5] distances for our ordinations. Using a single rarefaction to 1000 sequences, we visualized distances using non-metric multidimensional scaling (NMDS) plots in three dimensions. We then used adonis [1] to test the significance of the separation between sample types and locations in the full multidimensional space implied by the distance matrices.

We used bipartite graphs to help visualize associations between particular fungal OTUs and the three ant species. Since part of our interest in these fungal communities is to identify fungal taxa that may have ecological relevance, we focused on relatively abundant taxa by applying several filters to our OTU table. Although we cannot rule out apparently rare OTUs playing an important role, this would require, for example, sequencing biases to have reduced the apparent abundance of a common fungus, or for a rare fungus to have unusually large effects. We therefore chose to focus on the more abundant taxa. We included only OTUs with > 100 total sequences, and that were present in $\geqslant 2$ samples from the same ant species. These OTUs together comprised more than 90% of the rarefied dataset. Instances where an OTU contributed $\leqslant 10$ sequences to a sample were omitted from the graph for clarity (but an OTU may still be shown on the graph where it contributed > 10 sequences to other samples).

Data analysis – alate samples

To examine the alate samples, we combined OTU data from leaf, domatium and alate samples, and then rarefied to 300 sequences across the full dataset. The lower rarefaction threshold allowed us to include additional alate samples (5 alate samples had less than the 1000 sequence cutoff we used for the leaf and domatium samples), while still ensuring that the rarefied dataset was adequately representative.

We assessed alpha diversity of the alate samples using rarefaction curves and permutational statistical tests as described above for the domatium and leaf samples.

To gauge the possibility that alates might contribute to the domatium communities by vectoring fungi, we initially attempted to use distance-based ordination methods to assess the similarity of the alate samples to the domatium and leaf samples. However, distance ordination methods produced unhelpful representations of the alate data, with the low diversity alate samples giving rise to large intra- and intergroup distances. We therefore conducted an additional analysis by pooling the CN alate samples and determining the proportion of sequences in each ant-occupied Kenyan domatium that was found among the pooled alates; we performed an equivalent analysis for the pooled TP alate samples. This pooling approach allowed for the possibility that individual alates only take or contribute a fraction of the fungal community, and also helps overcome the limitations of applying our extraction and amplification methods to the tiny alate samples. Our opportunistic sampling did not allow such pooling at a finer scale (e.g. within a tree or within an ant colony).

References

[1] Marti J. Anderson. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46, 2001.

[2] Adrian Baddeley and Rolf Turner. spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005.

[3] E. Bent and D. L. Taylor. Direct amplification of DNA from fresh and preserved ectomycorrhizal root tips. *Journal of Microbiological Methods*, 80(2):206–208, 2010.

[4] R. Blaalid, S. Kumar, R. H. Nilsson, K. Abarenkov, P. M. Kirk, and H. Kauserud. ITS1 versus ITS2 as DNA metabarcodes for fungi. *Molecular Ecology Resources*, 13(2):218–224, 2013.

[5] J. Roger Bray and J. T. Curtis. An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27(4):325–349, 1957.

[6] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Tumbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, 2010.

[7] C.F. Dormann, B. Gruber, and J. Fruend. Introducing the bipartite package: analysing ecological networks. *R News*, 8(2):8–11, 2008.

[8] R. C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, 2010.

[9] M. Gardes and T. D. Bruns. ITS primers with enhanced specificity for basidiomycetes – application to the identification of mycorrhizae and rusts. *Molecular Ecology*, 2(2):113–118, 1993.

[10] J. B. Hughes and J. J. Hellmann. The application of rarefaction techniques to molecular inventories of microbial diversity. *Methods in Enzymology*, 397:292–308, 2005.

[11] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[12] C. Lozupone and R. Knight. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71(12):8228–35, 2005.

[13] P. J. McMurdie and S. Holmes. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Computational Biology*, 10(4):e1003531, 2014.

[14] Paul J. McMurdie. biom: an interface package (beta) for the BIOM file format, 2014.

[15] Erich Neuwirth. RColorBrewer: ColorBrewer palette: community ecology package, 2011.

[16] R. H. Nilsson, V. Veldre, M. Hartmann, M. Unterseher, A. Amend, J. Bergsten, E. Kristiansson, M. Ryberg, A. Jumpponen, and K. Abarenkov. An open source software package for automated extraction of ITS1 and ITS2 from fungal ITS sequences for use in high-throughput community assays and molecular ecology. *Fungal Ecology*, 3(4):284–287, 2010.

[17] Jari Oksanen, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, and Helene Wagner. vegan: community ecology package, 2013.

[18] T. M. Porter and G. B. Golding. Are similarity- or phylogeny-based methods more appropriate for classifying internal transcribed spacer (ITS) metagenomic amplicons? *The New Phytologist*, 192(3):775–82, 2011.

[19] C. Quince, A. Lanzen, T. P. Curtis, R. J. Davenport, N. Hall, I. M. Head, L. F. Read, and W. T. Sloan. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods*, 6(9):639–641, 2009.

[20] R Core Team. R: a language and environment for statistical computing, 2014.

[21] Karline Soetaert. shape: Functions for plotting graphical shapes, colors, 2014.

[22] Thorvald Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. *Biologiske Skrifter*, 5(4):1–34, 1948.

[23] Y. J. Sun, Y. P. Cai, V. Mai, W. Farmerie, F. H. Yu, J. A. Li, and S. Goodison. Advanced computational algorithms for microbial community analysis using massive 16S rRNA sequence data. *Nucleic Acids Research*, 38(22), 2010.

[24] T.J. White, T. Bruns, S. Lee, and J.W. Taylor. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In Michael A. Innis, D.H. Gelfand, J.J. Sninsky, and T. J. White, editors, *PCR Protocols : A Guide to Methods and Applications*, pages 315–322. Academic Press, San Diego, 1990.

[25] Hadley Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29, 2011.