OPEN ACCESS
phytobiomesjournal.org
Phytobiomes Journal
A Transdisciplinary Journal of Sustainable Plant Productivity

RESEARCH

e-Xtra*

Using the Microbiome Amplification Preference Tool (MAPT) to Reveal *Medicago sativa*-Associated Eukaryotic Microbes

Katherine Moccia,[1] Spiridon Papoulis,[1] Andrew Willems,[2] Zachary Marion,[3] James A. Fordyce,[4] and Sarah L. Lebeis[1,†]

[1] Department of Microbiology, University of Tennessee, Knoxville, TN, 37996, U.S.A.
[2] Department of Genome Science and Technology, University of Tennessee, Knoxville, TN, 37996-0840, U.S.A.
[3] School of Biology, University of Canterbury, Christchurch, 8041, New Zealand
[4] Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN, 37996-1610, U.S.A.

Accepted for publication 5 August 2020.

ABSTRACT

Although our understanding of the microbial diversity found within a given system expands as amplicon sequencing improves, technical aspects still drastically affect which members can be detected. Compared with prokaryotic members, the eukaryotic microorganisms associated with a host are understudied due to their underrepresentation in ribosomal databases, lower abundance compared with bacterial sequences, and higher ribosomal gene identity to their eukaryotic host. Peptide nucleic acid (PNA) blockers are often designed to reduce amplification of host DNA. Here we present a tool for PNA design called the Microbiome Amplification Preference Tool (MAPT). We examine the effectiveness of a PNA designed to block genomic *Medicago sativa* DNA (gPNA) compared with unrelated surrounding plants from the same location. We applied mitochondrial PNA and plastid PNA to block the majority of DNA from plant mitochondria and plastid 16S ribosomal RNA genes, as well as the novel gPNA. Until now, amplifying both eukaryotic and prokaryotic reads using 515F-Y and 926R has not been applied to a host. We investigate the efficacy of this gPNA using three approaches: (i) in silico prediction of blocking potential in MAPT, (ii) amplicon sequencing with and without the addition of PNAs, and (iii) comparison with cultured fungal representatives. When gPNA is added during amplicon library preparation, the diversity of unique eukaryotic amplicon sequence variants present in *M. sativa* increases. We provide a layered examination of the costs and benefits of using PNAs during sequencing. The application of MAPT enables scientists to design PNAs specifically to enable capturing greater diversity in their system.

Keywords: endophytes, microorganism, molecular biology, mycology, plants

Revealing the full microbial diversity of any environment is challenging. Although the isolation of novel organisms is essential to the core principles of microbiology, culture-dependent methods only provide a partial look at the overall diversity present on the planet, even in highly culturable systems such as plants (Bai et al. 2015; Carini 2019; Lloyd et al. 2018). At our current pace of 600 to 700 newly cultured microbial species per year, some scientists estimate it will take more than 1,000 years for all microorganisms to be cultured (Rosselló-Móra 2012; Yarza et al. 2014). Culture-independent methods such as amplicon sequencing introduce unintentional biases that also limit the ability to capture the true microbial diversity of any environment through the choice of primers, DNA extraction protocol, and amplicon library preparation (Fitzpatrick et al. 2018; Kiss 2012; Kovács et al. 2011; Lundberg et al. 2013; Nilsson et al. 2019; Parada et al. 2016; Sakai and Ikenaga 2013; Schoch et al. 2012; Terahara et al. 2011).

Definition of the eukaryotic members of microbiomes is widely performed by internally transcribed spacer (ITS) amplification, which primarily captures fungi (Schoch et al. 2012). Although ITS is the commonly accepted taxonomic identification for fungi, it has documented limitations, including taxonomically distinct copies within a single genome and low phylogenetic resolution (Kiss 2012; Kovács et al. 2011; Nilsson et al. 2008; Schoch et al. 2012). To

†Corresponding author: S. L. Lebeis; slebeis@vols.utk.edu

Author contributions: Z.M. performed the field sample collection for plants and arthropods. K.M. performed neighboring plant identification. K.M. and S.L.L. selected primers, designed the PNA experiments, and wrote the manuscript. K.M. prepared the amplicon sequencing libraries and generated the fungal isolate collection. The gPNA was designed by K.M., S.P., and S.L.L. S.P. generated and wrote the description for the Microbiome Amplification Preference Tool (MAPT) and generated the neighboring plant phylogenetic tree. The bioinformatic analysis of the 18S rRNA gene amplicon sequencing was performed by K.M. and A.W. Statistical analysis was performed by K.M. and J.A.F. K.M. and S.P. generated the figures and tables.

Funding: This work is supported by the National Science Foundation grant DEB-1638922 to S. Lebeis and J. Fordyce.

*The *e-Xtra* logo stands for "electronic extra" and indicates that supplementary tables and supplementary figures are published online.

The author(s) declare no conflict of interest.

uncover wider eukaryotic membership present in host microbiomes, another sequencing approach is required. Using primers such as 515F-Y and 926R that amplify both the 16S and 18S ribosomal RNA (rRNA) genes, scientists can capture both prokaryotes and eukaryotes (Needham et al. 2018; Parada et al. 2016). However, the 18S rRNA amplicons produced by these primers are often excluded from standard analysis because paired-end reads are usually too short to produce reads that overlap (Needham et al. 2018). Recent bioinformatic developments now enable scientists to analyze these reads without overlap, allowing the recovery of eukaryotic and prokaryotic reads with a single primer set (Lee 2019; Needham et al. 2018). The study of eukaryotic members of host-associated microbiomes is clouded by the vast abundance of host DNA and the lack of microbial eukaryotic representatives in sequence databases (Bai et al. 2015; Fitzpatrick et al. 2018; Liu et al. 2019; Lundberg et al. 2013). Although there are primers specific to the 18S rRNA gene (Liu et al. 2019), the addition of eukaryotes to community composition profiles without losing information about the bacterial community in a single amplicon library has provided a more extensive view of marine microbial communities (Needham et al. 2018; Parada et al. 2016). Within the context of the plant microbiome, 515F-Y and 926R have never been used to intentionally isolate eukaryotic reads. Utilizing these primers would allow the capture of protists and oomycetes, which standard ITS primers were not designed to amplify (Schoch et al. 2012). Although multiple 16S rRNA gene peptide nucleic acids (PNAs) were designed to block DNA from plant organelles, to our knowledge, no PNA has been designed to bind to the 18S rRNA gene present in the plant genome (Fitzpatrick et al. 2018; Lefèvre et al. 2020; Lundberg et al. 2013). 18S rRNA gene PNAs have been introduced successfully in other hosts such as mosquitoes and shrimp (Belda et al. 2017; Liu et al. 2019). However, it remains unclear if this approach will be successful to detect eukaryotic members of a plant microbiome.

The development of PNAs that bind to host DNA to block PCR amplification and, thus, increase microbial sequencing reads has been used for decades (Belda et al. 2017; Lefèvre et al. 2020; Lundberg et al. 2013; Ørum et al. 1993; von Wintzingerode et al. 2000). The most widely used PNAs to reduce plant read contamination were designed specifically to block plant 16S rRNA genes within *Arabidopsis thaliana* mitochondria and plastids (mPNA and pPNA, respectively), although other plants were queried for exact matches after PNA design was completed (Lundberg et al. 2013). Although mPNA and pPNA have since been used quite broadly to block DNA from organelles in other plants, it does not inhibit all plant DNA equally, which was predicted in the initial article (Lundberg et al. 2013). In fact, recent studies suggest that the design and use of a PNA must be specific to each host organism for effective plant DNA binding and subsequent blocking of amplification (Fitzpatrick et al. 2018). Although the desire for more robust PNAs is present, a flexible and easy design tool is still required.

The Earth Microbiome Project highlighted that host-associated microbial communities are less diverse than their surrounding free-living microbial communities, in both the number and abundance of each unique sequence, with the plant corpus as one of the least diverse microbial environments (Thompson et al. 2017). Furthermore, a study using a combined approach of whole-genome shotgun sequencing and amplicon analysis found bacterial reads to be 90% of microbial reads within the *A. thaliana* leaf microbiome, leaving eukaryotic microbes only the remaining 10% (Regalado et al. 2020). Plant tissues differ in host contamination when sequencing, because aboveground green tissue is known to have higher concentrations of DNA than other plant regions (Arenz et al. 2015; de Souza et al. 2016). Due to this decreased microbial diversity as well as the high abundance of host DNA, host ribosomal gene amplification must be prevented in order to enable examination of plant-associated eukaryotic microbes. Ideally, a PNA designed for the 18S rRNA gene would not interfere with amplification of fungi, Peronosporomycetes (oomycetes), or Cercozoan protists, which are all crucial members of the soil, phyllosphere, and endosphere of plants (Berney et al. 2017; de Araujo et al. 2018; Di Lucca et al. 2013; Geisen 2016; Jaskowska et al. 2015; McGhee and McGhee 1979; Ploch et al. 2016; Schwelm et al. 2018).

Here, we present a PNA designer called the Microbiome Amplification Preference Tool (MAPT) that allows researchers to (i) download the sequences desired to be both blocked as well as amplified, (ii) align the DNA region amplified by selected primers, (iii) find the region with the least similarity, and (iv) identify which organisms are at risk for unintentional amplification blockage. This enables researchers to design their own PNA with greater ease and to predict which organisms might have reduced detection with the addition of PNAs during amplicon library preparation. MAPT can be used with any host or environmental microbial community with representatives present in the SILVA database or with any FASTA sequences. Our novel 18S rRNA gene PNA, which we refer to as gPNA, was tested in the *Medicago sativa* phyllosphere and enable the detection of eukaryotic members of a host–microbe system.

MATERIALS AND METHODS

Creating the MAPT and genomic PNA for *M. sativa*. MAPT (https://github.com/SEpapoulis/MAPT) is a python module utilizing the publicly available data in SILVA, a high-quality ribosomal RNA database (https://www.arb-silva.de), to streamline the process of PNA development. SILVA was used because it is a comprehensive database for all three domains of life, with over 9 million small-subunit rRNA sequences (Yilmaz et al. 2014). Upon initialization, SILVA FTP servers are automatically queued for download, where users can specify whether the 'parc', 'ref', or 'nr ref99' datasets should be queued. After download, a local database is compiled for index-based searches using SILVA accessions, where sequence indexes are automatically organized under a taxonomic tree for convenient and rapid taxonomic searches. Using the SILVA database is optional, and users can alternatively specify sequences by providing their own FASTA files.

To select our PNA sequence to prevent host amplification, we performed a multiple sequence alignment for all *M. sativa* 18S rRNA genes available on SILVA version 132 to generate a consensus sequence. We aligned the *M. sativa* 18S rRNA gene consensus to all fungal sequences, as well as those from Peronosporomycetes (oomycetes) and Cercozoa sequences in the SILVA database, using MAPT. We chose fungi as well as the protists Peronosporomycetes (oomycetes) and Cercozoa because all were found in prior sequencing efforts in plant eukaryotic microbiome studies (de Araujo et al. 2018; Ploch et al. 2016; Schwelm et al.; 2018). We note that Peronosporomycetes were reclassified but the term "oomycetes" is still commonly utilized; therefore, we include it within this study in parentheses for clarity (Dick et al. 1999; Slater et al. 2013). Our PNA design was based on the methodology of Lundberg et al. (2013). We aligned the primers 515F-Y and 926R to the full 18S rRNA genes to extract the expected region amplified by our primer pair. Sequences were fragmented in silico into k-mers of 9 to 12 bases in length and aligned to the *M. sativa* sequence. We measured the total number of mapped k-mers to a specific DNA region (Fig. 1A). Our PNA sequence is the complement of the target sequence to allow binding because PNAs can bind parallel or antiparallel (Soomets et al. 1999).

We chose the region with the lowest identity to fungi and the two protist taxa that also satisfied custom PNA oligo guidelines (PNA Bio). Briefly, the PNA guidelines advised that the sequence be (i) less than 50% overall purine bases with no purine stretches more than 6 bases, (ii) less than 35% overall guanine bases, (iii) without significant complementarity to reduce the likelihood of hairpins, and (iv) shorter than 30 bases in length. Our resulting sequence, which was 12 bp long, was sent to PNA Bio to be created and quality tested.

The core design to MAPT follows a protocol similar to previous PNA design strategies, with slight modifications (Lundberg et al. 2013). DNA sequences from potential community members are cut into k-mers, or k-mer-sized DNA fragments, and mapped to exact matches in the host DNA sequence. Users can specify primers for in silico simulated amplification of all sequences provided before k-mers are generated and mapped. We note that the efficacy of a PNA changes depending on the primer set used because different primers will result in different distributions of k-mers. This prediction capability in MAPT could improve sequencing results and allows users greater flexibility. Our in silico amplification does not support degenerate k-mer mapping because exact matches to primers are required to be considered for subsequent k-mer analysis. However, any discarded sequences are reported for clarity in PNA design. To ensure that the underlying algorithms of our module were operating as intended, we recapitulated the mPNA and pPNA sequence selection and alignments using the Greengenes 16S rRNA dataset that was used for the initial generation of mPNA and pPNA (Fig. 1B and C) (Lundberg et al. 2013). We note that Lundberg et al. (2013) mapped k-mer sizes separately whereas ours are mapped together, resulting in differential graphical representation of k-mer alignment (Fig. 1).

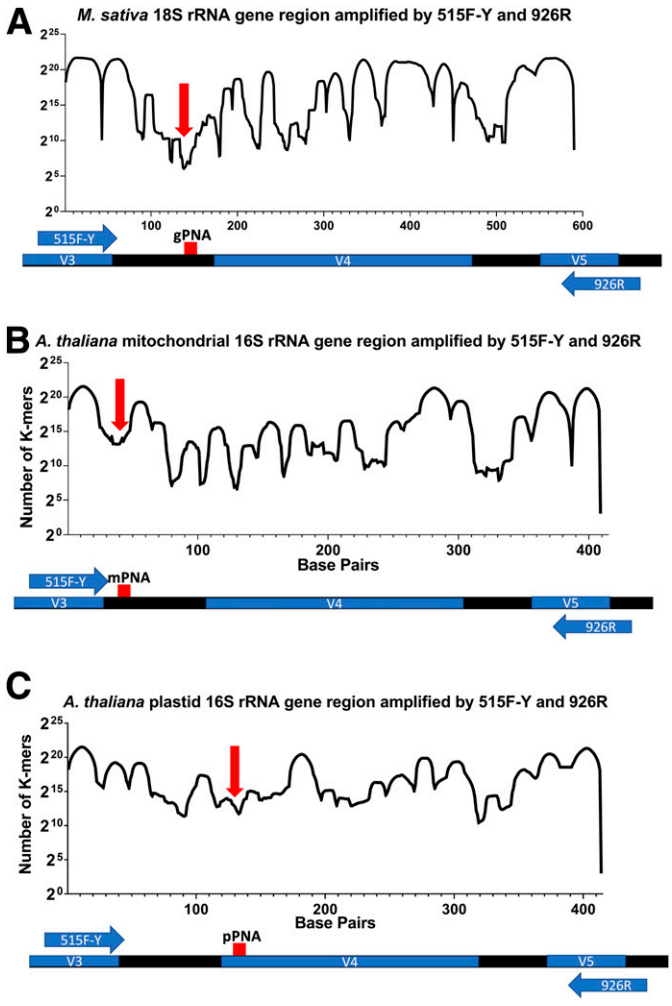

Fig. 1. Generation of novel genomic peptide nucleic acid (gPNA) and regeneration of mitochondrial and plastid PNA (mPNA and pPNA, respectively) demonstrates efficacy of MAPT. **A,** Mapping k-mers formed from fungal as well as Cercozoan and Peronosporomycetes protist reads to the *Medicago sativa* 18S ribosomal RNA (rRNA) gene region amplified by 515F and 926R. **B** and **C,** Regeneration of mPNA and pPNA from Lundberg et al. (2013). Schematic of where the PNAs (red boxes) are found on their respective genes. Horizontal blue arrows indicate the 515F and 926R primers used, with variable regions marked in blue, and red arrows indicate location of each PNA.

Field sample collection of feral *M. sativa* and neighboring plants. In all, 30 leaf samples from feral *M. sativa* and 10 samples from unrelated neighboring plants were removed aseptically and placed into Whirl-Pak bags (Consolidated Plastics). Feral plants are defined here as plants that were grown without human interference and cultivation. Leaf number per sample varied but all samples were standardized to 0.25 fresh weight and 0.04 dry weight in the lab. All samples were taken from Crystal Peak Park, Nevada at the coordinates 39.5134, –119.9955. Samples were placed on ice in the field then at 4°C until endophyte and epiphyte enrichments could be performed in the lab. To connect any arthropod 18S rRNA genes captured in our amplicon sequencing, insects were also collected using a sweep net with four sweeps for each *M. sativa* plant. All arthropods were then removed from the net using a manual aspirator and stored in alcohol until morphological identification could be performed. All arthropods were identified to the lowest taxonomic level that could be obtained, which was most often family level classification.

Endophyte and epiphyte sample enrichment. Phyllosphere epiphyte and endophyte sample separation was modified from Shade et al. (2013). There is no confirmation that all epiphyte material was removed from the leaves; thus, we refer to samples generated as epiphyte or endophyte enriched. Fresh plant tissue (0.25 g) from each sample was washed in 500 µl of 1× phosphate-buffered saline and 0.15% Tween 20 in 1-ml Eppendorf tubes. Samples were placed on ice and shaken for 20 min at 200 rpm, then sonicated for 5 min in a water bath (Branson). All liquid recovered from samples after sonication was classified as epiphyte-enriched material and ready for DNA extraction. The resulting plant material was flash frozen, then placed overnight in a FreezeZone Lyophilizer (Labconco). Lyophilized plant material was reweighed to 0.04 g dry weight and homogenized with approximately 20 sterile 0.7-mm garnet beads (Qiagen) in a Geno/Grinder 2010 (SPEX SamplePrep) for 30 s at 1,500 rpm. Once homogenized, samples were ready for DNA extraction as endophyte-enriched samples. All samples were extracted with the DNeasy Powersoil extraction kit (Qiagen) according to the suggested protocol. Replicate numbers ranged from 4 to 10 for all epiphyte and endophyte enrichments for both *M. sativa* and neighboring plants, for a combined total of 8 to 20 samples per PNA treatment (Supplementary Table S1).

Amplicon library preparation and sequencing. The PCR assays for the primer pair 515F-Y and 926R contained 2.5 µl of DNA (10 ng), 2.5 µl of 3-PNA mixture (mPNA to block the mitochondrial 16S rRNA gene, pPNA to block the plastid 16S rRNA gene, and gPNA to block the genomic 18S rRNA gene; 30 µM total), 12.5 µl of Hifi Hotstart Master Mix (KAPA Biosystems), and 5 µl of each primer (0.2 µM). The primer pair 515F-Y and 926R was used with the conditions defined by Parada et al. (2016), with an added 10-s addition before the primer annealing step to allow PNA binding in all PCR protocols used for sequencing. The conditions were as follows: 3 min at 95°C; then, 25

cycles of 95°C for 45 s, 78°C for 10 s, 50°C for 45 s, and 68°C for 90 s; and a final 68°C for 5 min.

Although we decided on the concentration of our gPNA addition to the PCR assays based on previous plant PNA design (Lundberg et al. 2013), we tested the ability of gPNA to block 18S rRNA gene amplification of heat-treated *M. sativa* DNA (*M. sativa* subsp. *sativa*, accession number 672758). When PNA concentration was increased from 0 to 30, 60, or even 100 μM in the PCR cocktail with universal primers 515F-Y and 926R using the above conditions, we still observe a band of the expected size of the 18S rRNA gene (Supplementary Fig. S1). Although previous studies used this method to select an appropriate concentration of PNA to add to their reactions (von Wintzingerode et al. 2000), we decided to not increase our concentration higher than suggested by Lundberg et al. (2013) because we observed that one of the two fungal isolates from *M. sativa* leaf tissue failed to amplify at the 100-μM PNA concentration (Supplementary Fig. S1). We used a heat treatment method to generate our plant tissue that was used previously to reduce the endophytic microbial population within *M. sativa* seed (Moccia et al. 2020). Briefly, we heated *M. sativa* seed for 30 min at 40°C, then rinsed for 1 min with 70% ethanol and 5 min of 10% freshly made bleach. At this temperature, the seed is sufficiently softened to allow for ethanol and bleach to kill seed endophytes. Seeds were then germinated on half-strength Murashige and Skoog germination agar with 1% sucrose (MP Biomedicals) in darkness for 2 days and light for 1 day. We refer to these seeds as heat-treated rather than sterile because there are a small number of microbial reads in these seedlings when sequenced but no colonies visible on this solid germination medium. The number of sequenced reads is significantly reduced from seeds that were not heat-treated (Moccia et al. 2020). DNA was extracted from the resulting seedlings with the DNeasy Plant mini kit (Qiagen).

ITS reactions contained 2.5 μl of DNA, 2.5 μl of 3-PNA mixture (16S mPNA, 16S pPNA, and 18S gPNA), 12.5 μl of Hifi Hotstart Master Mix (KAPA Biosystems), 0.83 μl of each of the six forward primers, and 2.5 μl of the two reverse primers. These primers amplified the ITS2 variable region. The conditions were as follows: 3 min at 95°C; then, 25 cycles of 95°C for 30 s, 78°C for 10 s, 55°C for 45 s, and 72°C for 30 s; and a final 72°C for 5 min.

For all primers, the samples with no PNA addition had sterile water added in lieu of PNA. The samples with two PNAs contained 2.5 μl of the 2-PNA mixture (mPNA and pPNA), with the same total concentration of the mPNA and pPNA as in the samples with all three PNAs (30 μM). Although the mPNA and pPNA were designed specifically for *A. thaliana*, both were predicted by Lundberg et al. (2013) to block *M. sativa* organelle amplification. Furthermore, we utilized mPNA and pPNA in a previous study to minimize *M. sativa* contamination in our 16S rRNA gene amplicon sequencing protocol (Moccia et al. 2020). Because we are using universal 16S rRNA primers to capture 18S rRNA genes, we decided to add mPNA and pPNA in addition to gPNA. All oligonucleotides used in amplicon sequencing are listed in Supplementary Table S2.

All samples were visualized on 1% agarose gels subsequently cleaned with Agencourt AMpure XP Beads (Beckman Coulter) according to the protocol from Illumina. Beads (20 μl) were added to each PCR sample and mixed by pipetting for 30 s per sample prior to a 10-min incubation. Beads with bound DNA were placed on a magnetic stand for approximately 5 min until the solution was clear and the supernatant was removed. Samples were washed with 80% fresh ethanol twice; then, 52.5 μl of 10 mM Tris HCl (Qiagen) was added to each sample. Tris HCl was mixed by pipetting for 30 s per sample. Samples were vortexed at 1,800 rpm, then incubated for 5 min on the magnetic bead stand. Cleaned DNA (49 to 50 μl) was pipetted off, leaving all magnetic beads adhered to the stand.

All amplicons received the same index PCR assays. Index PCR assays contained 5 μl of DNA, 5 μl of Nextera XT Index Forward Primer (Illumina), 5 μl of Nextera XT Index Reverse Primer (Illumina), 25 μl of KAPA Hifi Hotstart ReadyMix (KAPA Biosystems), 10 μl of PCR-grade water, and 2.5 μl of 3-PNA mixture. For the first reaction, the 2-PNA and 0-PNA samples alternatively contained the 2-PNA mixture and sterile water substitutions. The PCR protocol is as follows: 95°C for 3 min; then, eight cycles of 95°C for 30 s, 78°C for 10 s, 55° for 30 s, and 72°C for 30 s; and a final 72°C for 5 min. All samples were again visualized on 1% agarose gels, then cleaned with Agencourt AMpure XP Beads (Beckman Coulter) according to the protocol above, with only the modification of 56 μl of beads to 50 μl of PCR product and a final elution volume of 27.5 μl of Tris HCl to end the cleaning process with 24 to 25 μl of DNA. All samples were quantified using a Nanodrop 2000 (Thermo Fisher), then pooled in sets of eight based on Nanodrop results for approximately 500 ng/pool. Once pooled, samples were submitted to the University of Tennessee Genomics Core for analysis a Bioanalyzer High Sensitivity Chip (Agilent Technologies). Samples using 515F-Y and 926R primers were run on the Pippin Prep (Sage Science) to remove small (<80 bp) fragments on a 1.5% agar gel with the ranges for collection set at 525 to 875 bp. ITS samples did not require Pippin Prep because there were no small base pair fragments visible with the bioanalyzer. Pooled samples were cleaned once more with magnetic beads prior to sequencing with the same protocol as above. All samples were sequenced using the version 3, 600-cycle (2 × 300) kit on the Illumina MiSeq platform. Sequences have been submitted to the European Nucleotide Archive (ENA) at the European Molecular Biology Laboratory (EMBL) under the title "Eukaryotic Members of the Plant *Medicago Sativa*". These sequences can be found at under the primary accession PRJEB36800.

Separation of 18S rRNA gene amplicon reads. Because the 18S rRNA gene region amplified by 515F-Y and 926R is too long to overlap on a 2 × 300 paired-end sequencing run, additional bioinformatic steps were required to recover these reads. The protocol for separation of eukaryotic amplicon reads was from Happy Belly Bioinformatics (Lee 2019). Briefly, we downloaded and modified the Protist Ribosomal Reference (PR2) database (Guillou et al. 2013). After modifying the PR2 database by formatting the FASTQ files, we used the NCBI's Magic-BLAST application (Boratyn et al. 2019) to create a custom database. Following creation of the database, the 515F-Y and 926R primers were trimmed from all samples using the BBDuk tool (Joint Genome Institute) and all reads shorter than 250 bp were filtered out because these were likely to be bacterial sequences. The remaining samples were blasted using Magic-BLAST. Both forward and reverse reads were filtered with the requirements that >35% of the query sequence aligned within the database at >90% identity. If only a forward or reverse read passed the quality threshold, then it and its paired read were discarded. We then took the original fastq.gz files and the output from the Magic-BLAST step to split the reads of the fastq.gz files into four files. These files contained the forward and reverse reads for 16S and 18S rRNA gene reads. Because our recovery of 16S rRNA sequences was too low to allow comparisons between samples, we did not further analyze it. The 18S rRNA gene reads were processed in R using the DADA2 R package, version 1.10 (Callahan et al. 2016). For the 515F-Y and 926R primers, we captured 1,478,875 paired-end 18S rRNA gene reads in 77 total samples. There was a median of 10,748 18S rRNA gene reads per sample. There were 1,434 total 18S rRNA gene amplicon sequence variants (ASVs). We rarefied to 1,962 reads to perform all statistical analysis on rarefied data sets but, because abundance varies so much for 18S rRNA genes and rarefaction is a still debated technique, we

only rarefy for statistical analysis (de Vargas et al. 2015; McMurdie and Holmes 2014). Figures using rarefied data sets are specified within their figure legends (Supplementary Figs. S6, S10, and S13). Statistical estimates of Shannon's Diversity were measured using the R package phyloseq (McMurdie and Holmes 2013). Richness was among the metrics obtained using the R package vegetarian to calculate Hill numbers (Hill 1973; Jost 2006, 2007). Because we did not observe significant differences between epiphyte and endophyte ASV richness (Supplementary Fig. S13) or the relative abundance of plant reads (Supplementary Fig. S10), we analyzed the endophyte and epiphyte samples together from each PNA set, resulting in a replicate range of between 8 and 20 (Supplementary Table S1). Because richness, also known as q = 0, was the major finding with the addition of the gPNA, it has the most pertinent metric for evaluating sample differences.

Isolation of fungal collection from *M. sativa* samples. Plant material was collected at the coordinates 39.5102, –119.9952 in the Great Basin, located 0.5 miles from the original site where the *M. sativa* endophyte and epiphyte samples for DNA sequencing were collected. To isolate epiphyte samples, leaf and flower imprints were made on to the following media: lysogeny broth (LB) nutrients, 1/10 LB nutrients, 1/10 LB nutrients with 1% humic acid, 1/10 LB nutrients with 10% methanol, 1/5 dilution of King's B, MacConkey, and potato dextrose agar. To enrich for endophytes, leaves and flowers were surface sterilized with 10% household bleach and 0.01% Triton X-100 treatment. After 10 min submerged in bleach, leaves were washed with sterile distilled water. A solution of 2.5% sodium thiosulfate for 5 min neutralized the bleach. *M. sativa* leaves and flowers were washed twice more with sterile water. Approximately 20 sterile, 0.7-mm garnet beads (Qiagen) were added to the tubes in order to ensure sample homogenization on a GenoGrinder for 5 min at 1,500 rpm (SPEX SamplePrep). Homogenized endophyte-enriched samples were plated on the same media as the epiphyte-enriched samples and plates were incubated at 28°C for 2 weeks. Individual fungal hyphae were isolated using a dissecting microscope for visualization. DNA from all isolates was extracted using DNeasy Ultraclean Microbial Kit (Qiagen) and amplified with the ITS4 and ITS9 primer sets for fungi (Supplementary Table S3). Samples were cleaned with the QIAquick PCR Purification Kit (Qiagen) according to the protocol and submitted to University of Tennessee DNA Genomics Core for Sanger Capillary Sequencing.

Identification of neighboring plants. To interpret the *M. sativa* results with those from their neighboring plant samples, we attempted to classify all neighboring plants to genus level using molecular techniques. Our goal in using neighboring plants was to determine how well the gPNA blocks amplification of general plant DNA in comparison with the *M. sativa* samples, which it was specifically generated to bind. Even when PNAs are generated for a specific host, they are often applied to a variety of genetically similar hosts; thus, we included an examination of assorted plant material along with a phylogenetic tree comparing the alignment of gPNA across land plants (Fitzpatrick et al. 2018; Lundberg et al. 2013) (Supplementary Fig. S11). Because plant scientists do not agree on one primer set for identifying plant taxonomy, neighboring plants in the same field as feral *M. sativa* were identified using three common primer sets for rbcL, ITS2, and trnH-psbA (Supplementary Table S3). These primers are commonly used in combination with each other to identify plants from sequences (Fazekas et al. 2008; Li et al. 2015; Lledo et al. 1998; Hollingsworth et al. 2011; Stanford et al. 2000). DNA from the endophyte-enriched samples was used for plant identification PCRs (Supplementary Table S4). The PCR protocol was the same for rbcL and trnH-psbA: an initial 95°C step for 3 min; followed by 34 cycles of 95°C for 30 s, 57°C for 30 s, and 72°C for 1 min; and a final extension 72°C for 5 min. ITS2 had the same protocol except for a 53°C annealing temperature. As with the fungal isolate collection, samples were cleaned with the QIAquick PCR Purification Kit (Qiagen) according to the protocol and submitted to University of Tennessee DNA Genomics Core for Sanger Capillary Sequencing. Neighboring plant taxonomic identification was confirmed when at least two of the three genetic markers aligned to the same genus or genus and species (Supplementary Tables S4 and S5). We note that taxonomy results are based only on the molecular tools above and have not been confirmed via additional field collections.

Although the majority of the neighboring plants did not have representative 18S rRNA sequences in SILVA, the genera *Chamaenerion* and *Grindelia* did. Alignment revealed that *Chamaenerion* spp. did align completely with the gPNA, suggesting that it would be able to be blocked. However, *Grindelia* spp. did not contain a complementary sequence to the gPNA, suggesting that the gPNA would not block *Grindelia* 18S rRNA sequences. *Grindelia* spp. were the most isolated neighboring plant, with three samples identified. *Grindelia* spp. belong to the family Asteraceae. Although all other neighboring plants did not have a representative 18S rRNA sequence in SILVA, we know they also belong to the family Asteraceae (Supplementary Table S4). To visualize the gPNA alignments in a phylogenetic context, all aligned Magnoliophyta (land plant) sequences were downloaded from SILVA via MAPT (Supplementary Fig. S11). Alignment positions were masked if 30% of sequences contained a gap at a respective position. The masked multiple sequence alignment was then used to build a tree of Magnoliophyta using Fasttree 2.1 with a generalized time-reversable (-gtr) option (Price et al. 2010). MAPT was used to find the max gPNA k-mer in each Magnoliophyta sequence and ete3 was used to annotate the newick tree file with the k-mer data from MAPT (Huerta-Cepas et al. 2016). The tree was then uploaded and visualized with iTOL (Letunic and Bork 2019).

RESULTS

Design of a novel PNA to prevent host 18S rRNA gene amplification. To test the efficiency of our novel gPNA designed with MAPT, we decided to examine the eukaryotic microbial community associated with our selected *M. sativa* host using universal 515F-Y and 926R primers. Although the use of universal primers to capture eukaryotic reads has revealed the array of microbial eukaryotes in a marine environment (Needham et al. 2018), this has not been explored in plants previously due to the low abundance of eukaryotic microbial reads compared with host or bacterial reads (Regalado et al. 2020). To select potential gPNA sequences with minimal interference during eukaryotic microbiota amplification, the region of *M. sativa* predicted to be amplified by the 515F-Y and 926R primer set was aligned in MAPT with the 18S rRNA gene sequences present in SILVA assigned to all fungal as well as two protist taxa (Cercozoa and Peronosporomycetes) (Fig. 1A). This alignment revealed multiple regions of dissimilarity between *M. sativa* and eukaryotic microbial 18S rRNA gene sequences (Fig. 1A). The overall abundance and diversity of the k-mers were noted to account for the most sequenced organisms (Fig. 1A), which will have more representatives within the SILVA database. The gPNA sequence was created from the region with the least similarity to the k-mers, which also satisfied the PNA creation guidelines (see Materials and Methods). To test MAPT, the prediction of efficiency and specificity was also performed on *A. thaliana* 16S rRNA gene consensus sequences from plastids and mitochondria to recapitulate the pPNA and mPNA sequences previously published. MAPT was able to identify the location of

mPNA and pPNA, as well as provide the degree of similarity between potential bacterial microbiome members and the *A. thaliana* sequence (Fig. 1B and C) (Lundberg et al. 2013). We decided to subsequently investigate any potential bias against eukaryotic microbes associated with our new gPNA using another feature of MAPT.

Testing biases of PNA in silico. To predict the microbial eukaryotic members whose amplification might be blocked by our gPNA, we used the 18S rRNA gene region amplified by the primers 515F-Y and 926R to perform k-mer analysis of fungal sequences in SILVA (Supplementary Fig. S2), as well as total microbial eukaryotes (Supplementary Fig. S3). For fungal sequences, k-mer sizes ranged from 5 to 12, because our gPNA is 12 bases long. We found no fungal sequences in SILVA that aligned to k-mer sizes 9 to 12, and only 0.55 and 2.75% of the sequences aligned to sizes 7 and 8, respectively, suggesting that the gPNA would not have a high likelihood of blocking fungal amplicons (Supplementary Fig. S2A). We also included the sum of the k-mers that aligned to each organism with the size of k-mers weighted proportionally. For example, size 5 k-mers are weighted less than size 9 k-mers. We refer to this sum as the k-mer score (Supplementary Fig. S2B). Using these two metrics, we further examined the organisms with the highest k-mer scores, which included all organisms with the 8-k-mer length matches in Supplementary Fig. S2B (red bars). There were only 17 organisms from 13 genera with homology to the gPNA sequence (Supplementary Fig. S2C). Of these 13 genera, there was only 1 genus, *Absidia*, which was represented in our fungal isolate collection from *M. sativa* tissue. When we used DNA from the *Absidia* strain we isolated to determine whether gPNA prevented its 18S rRNA gene amplification, we observed no blockage (Supplementary Fig. S1, fungal isolate on the left), suggesting that our isolate is not among the organisms predicted by MAPT in Supplementary Fig. S2C. This result is consistent with amplicon sequencing results from samples that include the gPNA that contained reads for *Absidia* spp. present in the phyla Mucoromycota in Figure 2. Therefore, we did not to observe amplification blocking of this microbial taxon with our gPNA reagent, even though it was predicted to have the highest binding potential. This makes sense, considering that the maximum similarity found in the k-mer analysis was 8 bp and, thus, only 75% of the total gPNA. We next examined total microbial eukaryotes.

Within the total microbial eukaryotes represented in the SILVA database, no sequences aligned to k-mer sizes 11 or 12 and less than 2% of the sequences aligned to k-mer sizes 8 to 10, suggesting that the gPNA is not likely to block microbial eukaryotic amplification (Supplementary Fig. S3A). Among the 110 matches with the highest k-mer score and, therefore, the highest likelihood of being blocked by our gPNA (Supplementary Fig. S3B, red bar), 109 were found in marine or freshwater systems while 1 was found in a plant system (Supplementary Fig. S3C). Furthermore, none of the sequences were within the taxa of Peronosporomycetes. Thus, we did not discover large numbers of previously characterized plant-relevant eukaryotic microbes that might be blocked by the gPNA during amplicon library preparation. However, because a large percentage of diversity within potential plant inoculum such as soil remains uncharacterized, we cannot exclude the possibility that important eukaryotic microbes were negatively influenced by the addition of our gPNA during amplicon library preparation. Overall, in the analysis of fungal and total microbial eukaryotic reads, we found no sequences that had more than 83% alignment to our gPNA, and less than 3% of sequences had more than 80% alignment. This suggests that our gPNA was sufficiently specific for our intended host *M. sativa* and does not appear to have high similarity to microbial eukaryotes (Supplementary Figs. S2 and S3). Furthermore, this analysis performed by MAPT provides researchers with a list of organisms present in the SILVA database that have a higher likelihood of being affected by the addition of the PNA to investigate if they are concerned about bias against particular taxa in their amplicon sequencing.

Microbial eukaryotic members captured by 18S rRNA gene sequencing. Upon amplicon sequencing with the 515F-Y and 926R primers in our plant samples, we were able to successfully detect microbial eukaryotic reads. In total, 1,434 unique eukaryotic ASVs were recovered from the *M. sativa* and neighboring plant samples. The majority of these ASVs were microbial eukaryotes (66.8%), with the remaining ASVs divided between total plant eukaryotic ASVs (30.1%), unclassified ASVs (2.5%), and spurious bacteria (0.6%) (Fig. 2A). A small number of bacterial ASVs were expected because they were seen in previous use of this pipeline (Lee 2019). Upon examining the microbial eukaryotic ASVs, the largest portions belonged to the phylum Cercozoa, with almost a quarter of the overall ASVs (24.46%). The fungal phylum with the most ASVs was Ascomycota, with 20.58% of ASVs (Fig. 2B). Other top phyla from fungal, protist, and animal kingdoms include Ciliophora, Chytridiomycota, Basidiomycota, and Arthropoda, with each accounting for approximately 10% of ASVs (Fig. 2B; Supplementary Fig. S4). From our feral *M. sativa* plants, we collected insects to connect arthropod 18S rRNA gene ASVs observed (Fig. 2B; Supplementary Table S6). When matching at rank order, three of the four orders identified traditionally matched 18S rRNA gene sequences: Hemiptera, Hymenoptera, and Thysanoptera. All three arthropod orders sequenced are known herbivorous (Archibald et al. 2018; Eliyahu et al. 2015; Weirauch and Schuh 2011). Our 18S rRNA sequencing also found other plant-related eukaryotic organisms as well, including Peronosporomycetes (oomycetes) and the closely related Hyphochytriomycetes. The majority of the sequences from Peronosporomycetes classified at the genus level as *Pythium*, a prominent plant pathogen (Supplementary Fig. S5) (Schwelm et al. 2018). Thus, the 18S rRNA gene sequencing captured a wide array of eukaryotic diversity.

Connecting 18S rRNA gene to ITS amplicon sequencing and fungal isolation representatives. To compare the ASVs detected by 18S rRNA gene sequencing results to the commonly used ITS sequencing, we used a mixture of ITS2 region primers that contained six forward and two reverse primers with frameshifts in order to increase overall diversity of the amplicons sequenced (Cregger et al. 2018; White et al. 1990). We observed that the ITS sequences contained two fungal phyla sequences (Ascomycota and Basidiomycota) while the 18S rRNA gene sequencing captured these phyla plus an additional six not detected by ITS sequencing (Fig. 2B). Although protists were captured within the *M. sativa* samples for ITS primers, none of these ASVs were resolved to taxonomic orders below the phyla level, unlike for the 18S rRNA sequencing, where the phyla Cercozoa was resolved to the families Cercomonadidae, Glissomonadida, Imbricatea, Phytomyxea, and Thecofilosea and the phyla Ciliophora resolved to the subphyla Intramacronucleata and Postciliodesmatophora (Fig. 2B; Supplementary Fig. S4).

We compared the families represented in our fungal culture collection with those captured in the 18S rRNA gene and ITS amplicon sequencing results to see which primers have more cultured representatives (Fig. 2). Five of our eight isolated fungal families matched to 18S rRNA gene sequences present—Cladosporiaceae, Cunninghamellaceae, Incertae Sedis, Ophiocordycipitaceae, and Pleosporaceae—whereas only Pleosporaceae was detected by ITS sequencing (Fig. 2B to D; Supplementary Fig. S6). This difference is not likely due to the use of separate primers because the ITS primers we used to identify each member of the fungal isolate collection and to sequence the plants were both part of the same

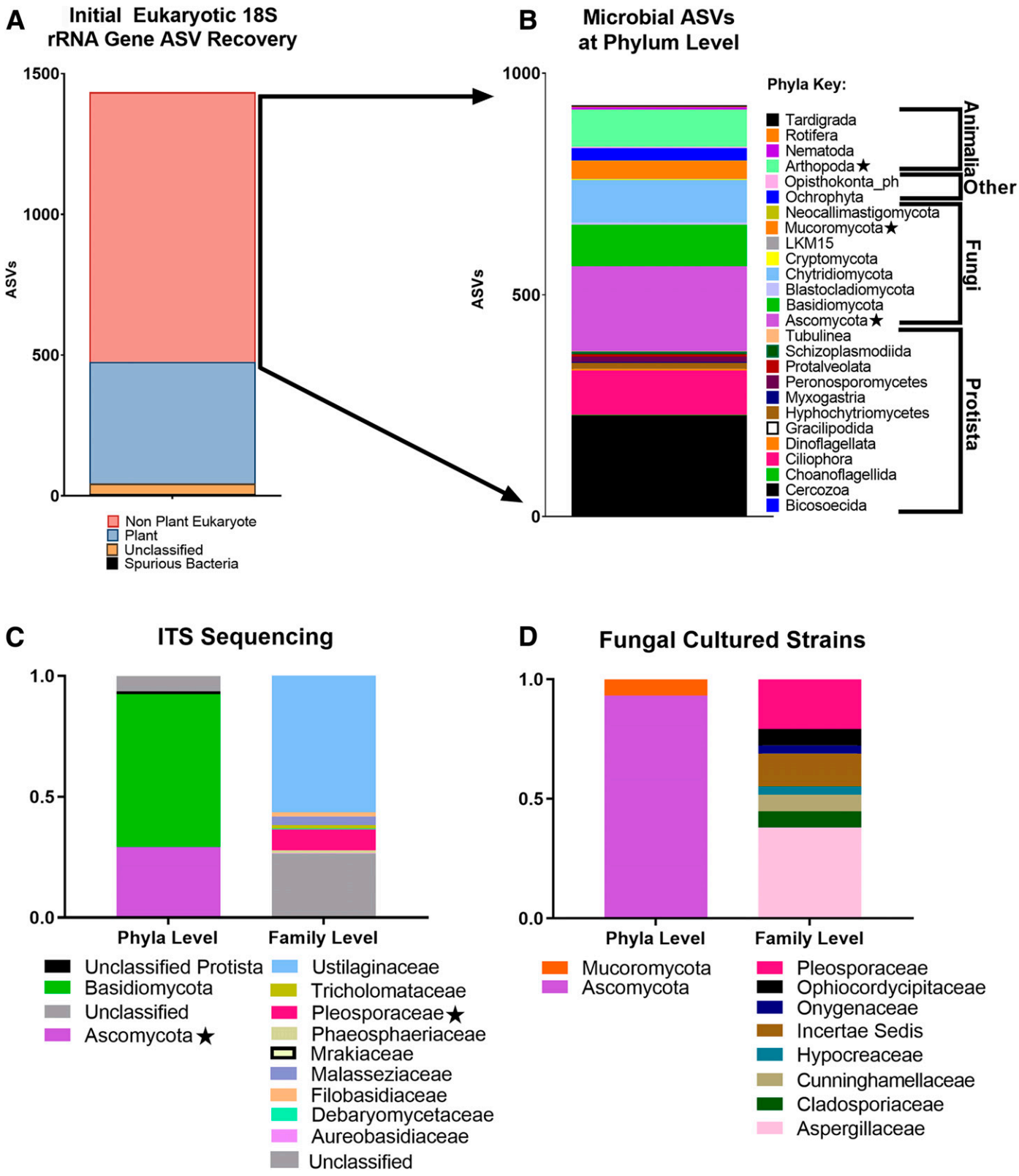

Fig. 2. 18S ribosomal RNA (rRNA) gene amplicon sequencing successfully captures a wide array of phyla across multiple kingdoms. **A,** Distribution of total amplicon sequence variants (ASVs) from all samples into main categories. **B,** Phylum-level distribution of microbial eukaryotic ASVs recovered from neighboring plants and *Medicago sativa* plants. "Other" indicates Orchophyta, which belongs to the Kingdom Chromista, and Opisthokonta, which belongs to both the Animal and Fungal Kingdoms. **C,** Internal transcribed spacer (ITS) amplicon sequencing results from *M. sativa* samples. **D,** Distribution of fungal culture collection generated from feral *M. sativa* tissue. Stars in B and C indicate isolated members of that taxonomic rank from *M. sativa* phyllosphere tissue.

variable region, ITS2. Although Incertae Sedis is used when the relation to other taxa is not known, the genus level of samples pertaining to the family Incertae Sedis corresponded with fungal isolates; thus, it was included. Three of these five families (i.e., Cladosporiaceae, Incertae Sedis, and Pleosporaceae) contained the most identified ASVs within the phyla Ascomycota. In all, 27 of the 29 fungal isolates in our culture collection belong to this phylum (Fig. 2D; Supplementary Fig. S6). Therefore, our 18S rRNA gene sequencing captured more ASVs with representatives in our culture collection than did ITS amplicon sequencing.

Influence of PNAs on 18S rRNA amplicon sequencing. The impact of the gPNA addition was measured by the changes in diversity of ASVs captured, as well as the number and relative abundance of *M. sativa* reads present. Due to the wide range in 18S rRNA genes per genome, abundance given by 18S rRNA gene sequencing is difficult to interpret (de Vargas et al. 2015; Needham et al. 2018). 18S rRNA gene copy can vary much more widely than the 16S rRNA gene, because one study demonstrated that the 18S rRNA gene can vary from 2 to 50,000 copies per genome while 16S rRNA copy number generally varies from 1 to 15 copies per genome (de Vargas et al. 2015; Kembel et al. 2012). Despite these copy number discrepancies, we use relative abundance of plant reads to see whether plant reads decreased with the addition of the gPNA because it is standard when measuring the design of a PNA (Supplementary Fig. S7) (Belda et al. 2017; Fitzpatrick et al. 2018; Lundberg et al. 2013).

The addition of the gPNA increased microbial eukaryotic ASV richness significantly within the *M. sativa* samples (Fig. 3A at $q = 0$) but not other metrics (Fig. 3), indicating that the ability to detect low-abundance or rare eukaryotic ASVs is increased by the addition of gPNA. To measure the diversity metrics of the rare versus high-abundance ASVs, we used Hill numbers, because rare community members are down weighted as q increases (Fig. 3A) (Jost 2006). Hill numbers at $q = 0$ are statistically equivalent to observed richness. The samples with all three PNA (i.e., gPNA, mPNA, and pPNA) were significantly increased in diversity when $q = 0$, although significantly decreased when $q = 2$ (Fig. 3A). Therefore, because rare ASVs are down weighted, the samples were dominated by plant reads and, at $q = 2$, our gPNA decreased the diversity of those abundant plant reads. After rarefaction, which can decrease detection of rare taxa (McMurdie and Holmes 2014), *M. sativa* samples sequenced appeared to be a highly selective system containing only four phyla (i.e., Arthropoda, Ascomycota, Basidiomycota, and Schizoplasmodiida) (Fig. 3C and D) although this would likely increase with larger samples. Although Arthropoda, Ascomycota, and Basidiomycota were present in all PNA combinations tested, Schizoplasmodiida was only present in the *M. sativa* samples with all three PNAs (Supplementary Fig. S8A). We confirmed that our difference in detecting various ASVs was not due to uneven replicate number between sample types by randomly subsampling to compare evenly across, suggesting that changes we saw were due to the addition of the gPNA (Supplementary Fig. S9). These data suggest that our gPNA is able to increase the diversity of the microbial eukaryotic ASVs detected.

To establish whether our gPNA is able to block host DNA amplification in other plants, we sampled from a selection of unrelated plants within the same field as our feral *M. sativa*. Differences in detected microbial eukaryotic ASVs between samples that contained no PNA, only mPNA and pPNA, or all three PNAs during library preparation did not translate into significant differences in diversity, although samples did follow the trends seen in *M. sativa* samples (Supplementary Fig. S10A and B). Interestingly, in these neighboring plant samples, six of the phyla missing in the samples with only mPNA and pPNA were present in both the 0-PNA and 3-PNA samples, suggesting that unintentional blockage by the pPNA and mPNA was potentially recovered with the addition of the gPNA samples (Supplementary Fig. S8B). Overall, neighboring plant samples displayed more diversity in the microbial eukaryotic community than in *M. sativa* samples (Supplementary Fig. S10C and D), possibly because there are six different genera of plants present, allowing for potentially a higher diversity of microorganisms to colonize within the various plants (Supplementary Tables S4 and S5). The variety of the neighboring plants could also account for why there is a wider range of diversity between samples because the gPNA could work better for some of these plant species than others. All neighboring plants are within the phyla Embryophyta, which comprises all land plants. We did not detect significant reduction of plant reads in these neighboring plant samples with the addition of the gPNA. Because our gPNA was designed specifically for *M. sativa* and neighboring plants were from a variety of backgrounds, these results are not surprising (Supplementary Fig. S10C; Supplementary Table S4).

We compared the 18S rRNA gene sequences for neighboring plant samples with representatives in SILVA with our gPNA sequence to determine the number of mismatches to the gPNA sequence. Of the two genera found within SILVA, *Chamaenerion* and *Grindelia*, the gPNA would require modifications to work for *Grindelia* but was predicted to block *Chamaenerion* (Supplementary Fig. S11). *Grindelia* is a part of the family Asteraceae, along with all other neighboring plants isolated, other than *Chamaenerion*. In comparison of 500 land plants, the family Asteraceae has previously been found to have the highest levels of plant plastid reads when using the pPNA (Fitzpatrick et al. 2018). We further created a phylogenetic tree to examine land plants in general (Supplementary Fig. S11). Green represents organisms that are predicted to be blocked by the gPNA because the sequence matches exactly to the gPNA based on k-mer alignment. It is evident that there is a great diversity in plant 18S rRNA sequences and that the gPNA would not function for all land plants. Organisms genetically similar to *Grindelia* spp. are not predicted to be blocked, offering an explanation for why the gPNA did not significantly increase neighboring plant samples because they are more phylogenetically related to *Grindelia* than *Chamaenerion* (Supplementary Fig. S11). Overall, the variance within the neighboring plants highlights the need for a host-specific PNA to be designed.

For *M. sativa* endophyte-enriched or epiphyte-enriched samples, we failed to detect a decrease in the relative abundance of plant reads with the addition of the gPNA during amplicon library preparation (Supplementary Fig. S7A), which corresponds to the limited diversity differences only in rare ASVs when $q = 0$ (Fig. 3A) and the lack of disappearance of an 18S rRNA band with the addition of the gPNA on heat-treated *M. sativa* plants (Supplementary Fig. S1). Interestingly, the proportion of microbial to plant reads did increase modestly for *M. sativa* samples amplified with all three PNAs in both endophyte and epiphyte samples (Supplementary Fig. S12). Samples dominated by host eukaryotic reads had been previously observed with the application of PNA, where host eukaryotic reads were a fold higher than their microbial eukaryotic reads (Belda et al. 2017). Overall, the small increase in microbial/plant read ratio supports the conclusion that our gPNA helps to increase the diversity of low-abundance ASVs, specifically in *M. sativa* samples, because the changes seen in read abundance for these rare ASVs would be small.

DISCUSSION

Regardless of the technical parameters, amplicon sequencing of the microbial community of any system comes with inherent

detection limitations. Here, we demonstrated that 515F-Y and 926R primers can amplify low-abundance eukaryotic reads in a host setting with MAPT enabling effective PNA design and specificity predictions (Fig. 1). The addition of our 18S gPNA increases the number and diversity of microbial eukaryotic ASVs detected specifically in *M. sativa* samples (Fig. 3). We also show that, overall, 18S rRNA sequencing is able to recover a more diverse number of fungal taxa than ITS sequences when we compare both to the members of our culture collection as well as in the diversity of fungal phyla sequenced (Fig. 2; Supplementary Fig. S6). Finally, we recovered additional ASVs that correspond to organisms such as protists and arthropods (Fig. 2; Supplementary Fig. S4). Although we did not detect a significant decrease the number of plant reads, our gPNA accomplished the overall goal of increasing the eukaryotic diversity captured in a eukaryotic host by revealing rare taxa (Fig. 3; Supplementary Fig. S4). This is consistent with a study that generated 18S rRNA gene PNA against mosquitoes to reveal malarial burden, where two PNAs were generated in different variable regions (Belda et al. 2017). In this article, the reduction in the host reads was small (less than 10%) in the PNA generated in variable region 9 (V9) of the rRNA gene. Both PNAs were tested at 0.75, 1.5, 3.75, and 7.5 μM and, whereas the V9 PNA worked most efficiently at 7.5, the other PNA, located in the V4 region, contained the maximum eukaryotic microbiota reads at 1.5 μM PNA. However, the slight decrease in host reads in the V9 PNA allowed for eukaryotic microbe richness to increase (Belda et al. 2017).

Design and implementation of a PNA can greatly reduce unwanted amplification, especially of 16S and 18S rRNA gene sequencing in host systems (Fitzpatrick et al. 2018; Lefèvre et al. 2020; Lundberg et al. 2013). However, inclusion of a PNA during amplicon library preparation must always be done with strict effort to minimize the unintentional blocking of other organisms and to document any bias that is observed with the addition of this reagent. MAPT is able to minimize the likelihood that a PNA will have unwanted blockage by aligning the host sequence to any taxonomic group desired. Because MAPT utilizes the researcher's chosen primer pair, it selects for the community amplified by those primers and, thus, enables higher specificity for the wide array of primers possible within sequencing protocols. MAPT also allows for a variety of design opportunities, from highly diverse environments to those where perhaps only a few genera are found. Furthermore, our tool is able to identify sequences that are at risk of being blocked because it can detect sequences with high similarity to the PNA in the reference database (Supplementary Figs. S2 and S3). We further advise scientists implementing PNAs into library procedures to add PNA-free controls within their samples to be sequenced to establish the bias introduced by a PNA on the overall community diversity.

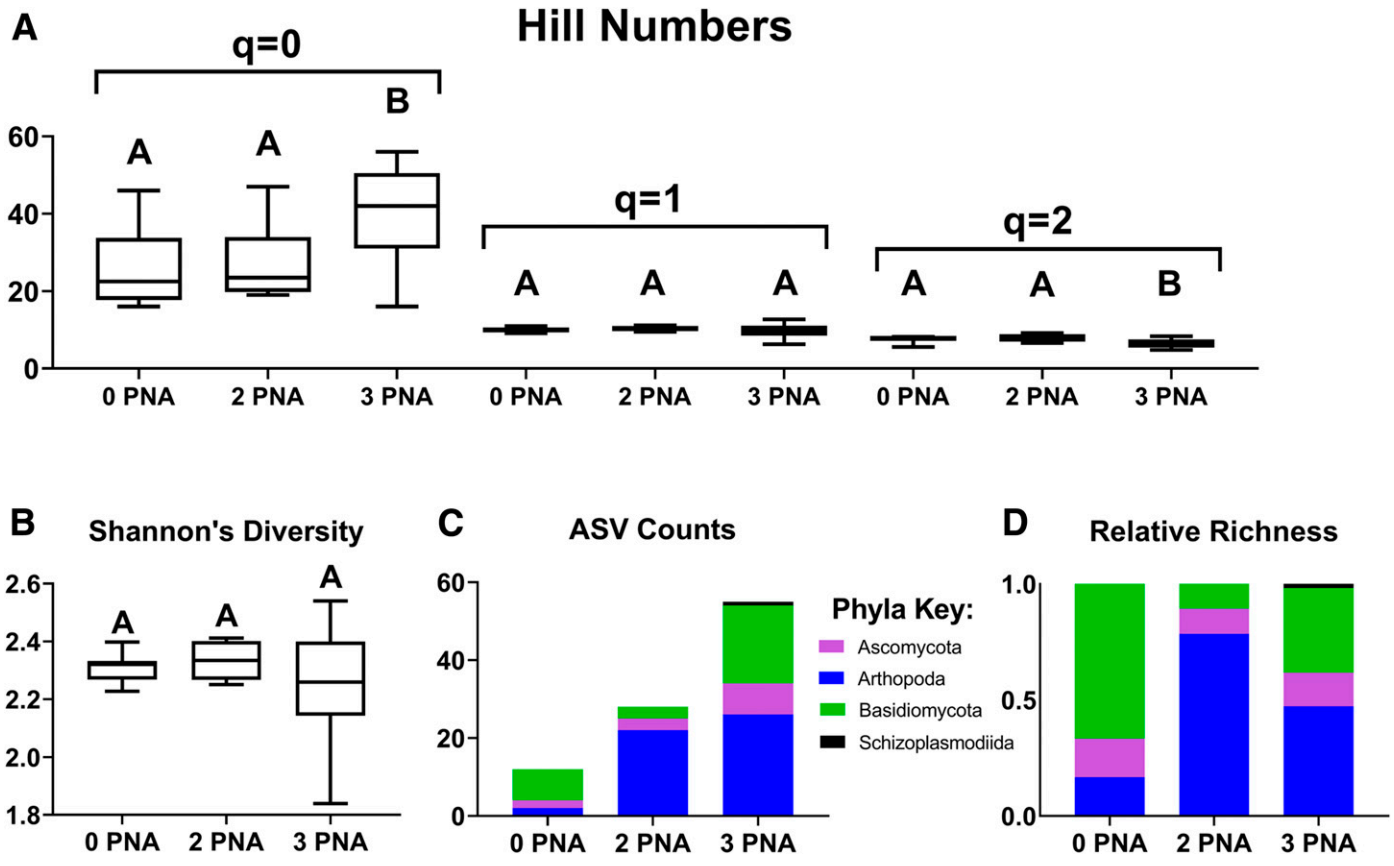

Fig. 3. Species richness in *Medicago sativa* rarefied samples increases with novel genomic peptide nucleic acid (gPNA) addition. Amplicon sequence variants (ASVs) present in *M. sativa* samples amplified with no PNA (0 PNA); only mitochondrial and plastid PNA (mPNA and pPNA, respectively) (2 PNA); or gPNA, mPNA, and pPNA (3 PNA) were analyzed for a variety of diversity metrics. **A,** Hill Numbers ($F_{2,31}$ = 6.841, 0.6988, and 8.175 for q = 0 to 3, respectively. At q = 0: *P* value for 0 and 2 PNA = 0.9756, for 0 and 3 PNA = 0.0123, for *P* value for 2 and 3 PNA = 0.013. At q = 1: *P* value = 0.5048. At q = 2: *P* value for 0 and 2 PNA = 0.7362, for 0 and 3 PNA = 0.0294, for *P* value for 2 and 3 PNA = 0.0019). **B,** Shannon's diversity ($F_{2,31}$ = 1.246, *P* value = 0.3016). **C,** Total ASV counts. **D,** Relative richness (A and B) significance was determined with an analysis of variance (ANOVA) with a post hoc Tukey's multiple comparison test, α = 0.05 for all comparisons. *P* values for multiple comparisons were only reported when ANOVA revealed significance. See Supplementary Table S1 for sample size.

Conclusion. The PNA that we designed cannot be applied universally for all plant species. We tested with neighboring plants because PNAs are frequently designed for one species of host but applied to phylogenetically similar organisms (Supplementary Figs. S10 and S11). In fact, the previously published pPNA has been shown to be less effective against species within the family Asteraceae (Fitzpatrick et al. 2018). The result of the neighboring plant sequencing further demonstrates the need for MAPT, because PNAs should be designed specifically for each host and MAPT enables this to be performed with greater ease. The neighboring plants sampled here belong to only two families, Asteraceae and Onagraceae; therefore, MAPT captures a small window of genetic variation within plants as a whole (Supplementary Table S4) (Royal Botanic Gardens and Missouri Botanical Garden 2013). Thus, the likelihood is minute that any PNA designed for a specific species of plant, such as the gPNA, would be able to be applied across all plant species. However, small base change modifications in PNAs, such as shown by Fitzpatrick et al. (2018), can allow scientists to take a PNA designed for one plant species and modify it slightly to be used for another. A small modification to our gPNA may enable it to be applied to another plant type. With the use of MAPT and the presentation of its application here, we hope to encourage scientists to design their own PNAs when undergoing amplicon sequencing with a novel host, in order to obtain the highest level of PNA efficacy and specificity. The design and implementation of PNAs across host-associated systems will help to reveal more members of the microbial community otherwise left not sequenced, leading to a more accurate depiction of the relevant and consistent members of host-associated microbiomes.

ACKNOWLEDGMENTS

We thank S. Herrera Paredes, J. Harrison, C. Nice, and M. Forister for their advice and interpretation of the results; G. Forister for the insect identification; and V. Brown and R. Murdoch for their generous support with Illumina sequencing and analysis.

LITERATURE CITED

Archibald, S. B., Rasnitsyn, A. P., Brothers, D. J., and Mathewes, R. W. 2018. Modernisation of the Hymenoptera: Ants, bees, wasps, and sawflies of the early Eocene Okanagan Highlands of western North America. Can. Entomol. 150:205-257.

Arenz, B. E., Schlatter, D. C., Bradeen, J. M., and Kinkel, L. L. 2015. Blocking primers reduce co-amplification of plant DNA when studying bacterial endophyte communities. J. Microbiol. Methods 117:1-3.

Bai, Y., Müller, D. B., Srinivas, G., Garrido-Oter, R., Potthoff, E., Rott, M., Dombrowski, N., Münch, P. C., Spaepen, S., Remus-Emsesrmann, M., Hüttel, B., McHardy, A. C., Vorholt, J. A., and Schulze-Lefert, P. 2015. Functional overlap of the Arabidopsis leaf and root microbiota. Nature 528: 364-369.

Belda, E., Coulibaly, B., Fofana, A., Beavogui, A. H., Traore, S. F., Gohl, D. M., Vernick, K. D., and Riehle, M. M. 2017. Preferential suppression of *Anopheles gambiae* host sequences allows detection of the mosquito eukaryotic microbiome. Sci. Rep. 7:3241.

Berney, C., Ciuprina, A., Bender, S., Brodie, J., Edgcomb, V., Kim, E., Rajan, J., Parfrey, L. W., Adl, S., Audic, S., Bass, D., Caron, D. A., Cochrane, G., Czech, L., Dunthorn, M., Geisen, S., Glöckner, F. O., Mahé, F., Quast, C., Kaye, J. Z., Simpson, A. G. B., Stamatakis, A., del Campo, J., Yilmaz, P., and de Vargas, C. 2017. *UniEuk*: Time to speak a common language in protistology! J. Eukaryot. Microbiol. 64:407-411.

Boratyn, G. M., Thierry-Mieg, J., Thierry-Mieg, D., Busby, B., and Madden, T. L. 2019. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. BMC Bioinf. 20:405.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods 13:581-583.

Carini, P. 2019. A "Cultural" Renaissance: Genomics breathes new life into an old craft. mSystems 4:e00092-19.

Cregger, M. A., Veach, A. M., Yang, Z. K., Crouch, M. J., Vilgalys, R., Tuskan, G. A., and Schadt, C. W. 2018. The *Populus* holobiont: Dissecting the effects of plant niches and genotype on the microbiome. Microbiome 6:31.

de Araujo, Mendes, A. S. F., L. W., Lemos, L. N., Antunes, J. E. L., Beserra, J. E. A., Jr., M., de Lyra, do C. C. P., Figueiredo, M. do V. B., de Almeida Lopes, Â. C., Gomes, R. L. F., Bezerra, W. M., Melo, V. M. M., de Araujo, F., and Geisen, S. 2018. Protist species richness and soil microbiome complexity increase towards climax vegetation in the Brazilian Cerrado. Commun. Biol. 1:138. https://www.nature.com/articles/s42003-018-0129-0

de Souza, R. S. C., Okura, V. K., Armanhi, J. S. L., Jorrín, B., Lozano, N., da Silva, M. J., González-Guerrero, M., de Araújo, L. M., Verza, N. C., Bagheri, H. C., Imperial, J., and Arruda, P. 2016. Unlocking the bacterial and fungal communities assemblages of sugarcane microbiome. Sci. Rep. 6: 28774.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horák, A., Jaillon, O., Lima-Mendez, G., Lukeš, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans Coordinators, Acinas, S. G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M. E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., and Karsenti, E. 2015. Eukaryotic plankton diversity in the sunlit ocean. Science 348:1261605.

Dick, M. W., Vick, M. C., Gibbings, J. G., Hedderson, T. A., and Lopez Lastra, C. C. 1999. 18S rDNA for species of Leptolegnia and other Peronosporomycetes: Justification for the subclass taxa Saprolegniomycetidae and Peronosporomycetidae and division of the Saprolegniaceae sensu lato into the Leptolegniaceae and Saprolegniaceae. Mycol. Res. 103:1119-1125.

Di Lucca, A. G. T., Trinidad Chipana, E. F., Talledo Albújar, M. J., Dávila Peralta, W., Montoya Piedra, Y. C., and Arévalo Zelada, J. L. 2013. Slow wilt: Another form of Marchitez in oil palm associated with trypanosomatids in Peru. Trop. Plant Pathol. 38:522-533.

Eliyahu, D., McCall, A. C., Lauck, M., Trakhtenbrot, A., and Bronstein, J. L. 2015. Minute pollinators: The role of thrips (Thysanoptera) as pollinators of pointleaf manzanita, *Arctostaphylos pungens* (Ericaceae). J. Pollinat. Ecol. 16:64-71. http://www.pollinationecology.org/index.php?journal=jpe&page=issue&op=view&path%5B%5D=47

Fazekas, A. J., Burgess, K. S., Kesanakurti, P. R., Graham, S. W., Newmaster, S. G., Husband, B. C., Percy, D. M., Hajibabaei, M., and Barrett, S. C. H. 2008. Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. PLoS One 3:e2802.

Fitzpatrick, C. R., Lu-Irving, P., Copeland, J., Guttman, D. S., Wang, P. W., Baltrus, D. A., Dlugosch, K. M., and Johnson, M. T. J. 2018. Chloroplast sequence variation and the efficacy of peptide nucleic acids for blocking host amplification in plant microbiome studies. Microbiome 6:144.

Geisen, S. 2016. The bacterial-fungal energy channel concept challenged by enormous functional versatility of soil protists. Soil Biol. Biochem. 102: 22-25.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H. C. F., Lara, E., Le Bescot, N., Logares, R., Mahé, F., Massana, R., Montresor, M., Morard, R., Not, F., Pawlowski, J., Probert, I., Sauvadet, A.-L., Siano, R., Stoeck, T., Vaulot, D., Zimmermann, P., and Christen, R. 2013. The Protist Ribosomal Reference database (PR^2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic Acids Res. 41: D597-D604.

Hill, M. O. 1973. Diversity and evenness: A unifying notation and its consequences. Ecology 54:427-432.

Hollingsworth, P. M., Graham, S. W., and Little, D. P. 2011. Choosing and using a plant DNA barcode. PLoS One 6:e19254.

Huerta-Cepas, J., Serra, F., and Bork, P. 2016. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. Mol. Biol. Evol. 33: 1635-1638.

Jaskowska, E., Butler, C., Preston, G., and Kelly, S. 2015. Phytomonas: Trypanosomatids adapted to plant environments. PLOS Pathog. 11:e1004484.

Jost, L. 2006. Entropy and diversity. Oikos 113:363-375.

Jost, L. 2007. Partitioning diversity into independent alpha and beta components. Ecology 88:2427-2439.

Kembel, S. W., Wu, M., Eisen, J. A., and Green, J. L. 2012. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. PLOS Comput. Biol. 8:e1002743.

Kiss, L. 2012. Limits of nuclear ribosomal DNA internal transcribed spacer (ITS) sequences as species barcodes for fungi. Proc. Natl. Acad. Sci. U.S.A. 109:E1811.

Kovács, G. M., Jankovics, T., and Kiss, L. 2011. Variation in the nrDNA ITS sequences of some powdery mildew species: Do routine molecular identification procedures hide valuable information? Eur. J. Plant Pathol. 131:135-141.

Lee, M. D. 2019. Happy Belly Bioinformatics: An open-source resource dedicated to helping biologists utilize bioinformatics. J. Open Source Educ. 2: 53.

Lefèvre, E., Gardner, C. M., and Gunsch, C. K. 2020. A novel PCR-clamping assay reducing plant host DNA amplification significantly improves prokaryotic endo-microbiome community characterization. FEMS Microbiol. Ecol. 96:fiaa110.

Letunic, I., and Bork, P. 2019. Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. Nucleic Acids Res. 47:W256-W259.

Li, X., Yang, Y., Henry, R. J., Rossetto, M., Wang, Y., and Chen, S. 2015. Plant DNA barcoding: From gene to genome. Biol Rev. 90:157-166.

Liu, C., Qi, R.-J., Jiang, J.-Z., Zhang, M.-Q., and Wang, J.-Y. 2019. Development of a blocking primer to inhibit the PCR amplification of the 18S rDNA sequences of *Litopenaeus vannamei* and its efficacy in *Crassostrea hongkongensis*. Front. Microbiol. 10:830.

Lledo, M. D., Crespo, M. B., Cameron, K. M., Fay, M. F., and Chase, M. W. 1998. Systematics of Plumbaginaceae based upon cladistic analysis of rbcL sequence data. Syst. Bot. 23:21.

Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J., and Crosby, L. 2018. Phylogenetically novel uncultured microbial cells dominate earth microbiomes. mSystems 3: e00055-18.

Lundberg, D. S., Yourstone, S., Mieczkowski, P., Jones, C. D., and Dangl, J. L. 2013. Practical innovations for high-throughput amplicon sequencing. Nat. Methods 10:999-1002.

McGhee, R. B., and McGhee, A. H. 1979. Biology and structure of *Phytomonas staheli* sp. n., a trypanosomatid located in sieve tubes of coconut and oil palms. J. Protozool. 26:348-351.

McMurdie, P. J., and Holmes, S. 2013. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8: e61217.

McMurdie, P. J., and Holmes, S. 2014. Waste not, want not: Why rarefying microbiome data is inadmissible. PLOS Comput. Biol. 10:e1003531.

Moccia, K., Willems, A., Papoulis, S., Flores, A., Forister, M. L., Fordyce, J. A., and Lebeis, S. L. 2020. Distinguishing nutrient-dependent plant driven bacterial colonization patterns in alfalfa. Environ. Microbiol. Rep. 12:70-77.

Needham, D. M., Fichot, E. B., Wang, E., Berdjeb, L., Cram, J. A., Fichot, C. G., and Fuhrman, J. A. 2018. Dynamics and interactions of highly resolved marine plankton via automated high-frequency sampling. ISME J. 12: 2417-2432.

Nilsson, R. H., Anslan, S., Bahram, M., Wurzbacher, C., Baldrian, P., and Tedersoo, L. 2019. Mycobiome diversity: High-throughput sequencing and identification of fungi. Nat. Rev. Microbiol. 17:95-109.

Nilsson, R. H., Kristiansson, E., Ryberg, M., Hallenberg, N., and 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. Evol. Bioinf. 2:193-201.

Ørum, H., Nielsen, P. E., Egholm, M., Berg, R. H., Buchardt, O., and Stanley, C. 1993. Single base pair mutation analysis by PNA directed PCR clamping. Nucleic Acids Res. 21:5332-5336.

Parada, A. E., Needham, D. M., and Fuhrman, J. A. 2016. Every base matters: Assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environ. Microbiol. 18: 1403-1414.

Ploch, S., Rose, L. E., Bass, D., and Bonkowski, M. 2016. High diversity revealed in leaf-associated protists (Rhizaria: Cercozoa) of Brassicaceae. J. Eukaryot. Microbiol. 63:635-641.

Price, M. N., Dehal, P. S., and Arkin, A. P. 2010. FastTree 2—Approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.

Regalado, J., Lundberg, D. S., Deusch, O., Kersten, S., Karasov, T., Poersch, K., Shirsekar, G., and Weigel, D. 2020. Combining whole-genome shotgun sequencing and rRNA gene amplicon analyses to improve detection of microbe–microbe interaction networks in plant leaves. ISME J. 14: 2116-2130.

Rosselló-Móra, R. 2012. Towards a taxonomy of *Bacteria* and *Archaea* based on interactive and cumulative data repositories. Environ. Microbiol. 14:318-334.

Royal Botanic Gardens and Missouri Botanical Garden. 2013. The Plant List, version 1.1. http://www.theplantlist.org/

Sakai, M., and Ikenaga, M. 2013. Application of peptide nucleic acid (PNA)-PCR clamping technique to investigate the community structures of rhizobacteria associated with plant roots. J. Microbiol. Methods 92:281-288.

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., Chen, W., and Fungal Barcoding Consortium. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc. Natl. Acad. Sci. U.S.A. 109:6241-6246.

Schwelm, A., Badstöber, J., Bulman, S., Desoignies, N., Etemadi, M., Falloon, R. E., Gachon, C. M. M., Legreve, A., Lukeš, J., Merz, U., Nenarokova, A., Strittmatter, M., Sullivan, B. K., and Neuhauser, S. 2018. Not in your usual Top 10: Protists that infect plants and algae. Mol. Plant Pathol. 19:1029-1044.

Shade, A., McManus, P. S., and Handelsman, J. 2013. Unexpected diversity during community succession in the apple flower microbiome. MBio 4: e00602-12.

Slater, B. J., McLoughlin, S., and Hilton, J. 2013. Peronosporomycetes (Oomycota) from a Middle Permian permineralised peat within the Bainmedart Coal Measures, Prince Charles Mountains, Antarctica. PLoS One 8:e70707.

Soomets, U., Hällbrink, M., and Langel, Ü. 1999. Antisense properties of peptide nucleic acids. Front. Biosci. 4:D782-D786.

Stanford, A. M., Harden, R., and Parks, C. R. 2000. Phylogeny and biogeography of *Juglans* (Juglandaceae) based on *matK* and ITS sequence data. Am. J. Bot. 87:872-882.

Terahara, T., Chow, S., Kurogi, H., Lee, S.-H., Tsukamoto, K., Mochioka, N., Tanaka, H., and Takeyama, H. 2011. Efficiency of peptide nucleic acid-directed PCR clamping and its application in the investigation of natural diets of the Japanese eel Leptocephali. PLoS One 6:e25715.

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., Prill, R. J., Tripathi, A., Gibbons, S. M., Ackermann, G., Navas-Molina, J. A., Janssen, S., Kopylova, E., Vázquez-Baeza, Y., González, A., Morton, J. T., Mirarab, S., Xu, Z. Z., Jiang, L., Haroon, M. F., Kanbar, J., Zhu, Q., Song, S. J., Kosciolek, T., Bokulich, N. A., Lefler, J., Brislawn, C. J., Humphrey, G., Owens, S. M., Hampton-Marcell, J., Berg-Lyons, D., McKenzie, V., Fierer, N., Fuhrman, J. A., Clauset, A., Stevens, R. L., Shade, A., Pollard, K. S., Goodwin, K. D., Jansson, J. K., Gilbert, J. A., Knight, R., and The Earth Microbiome Project Consortium. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457-463.

von Wintzingerode, F., Landt, O., Ehrlich, A., and Göbel, U. B. 2000. Peptide nucleic acid-mediated PCR clamping as a useful supplement in the determination of microbial diversity. Appl. Environ. Microbiol. 66:549-557.

Weirauch, C., and Schuh, R. T. 2011. Systematics and evolution of Heteroptera: 25 Years of progress. Annu. Rev. Entomol. 56:487-510.

White, T. J., Bruns, T., Lee, S., and Taylor, J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. Pages 315-322 in: PCR Protocols: A Guide to Methods and Applications. Academic Press, San Diego, CA.

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K.-H., Whitman, W. B., Euzéby, J., Amann, R., and Rosselló-Móra, R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat. Rev. Microbiol. 12:635-645.

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., and Glöckner, F. O. 2014. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res. 42:D643-D648.