Assessment by Matrix-Assisted Laser Desorption – Time of Flight Mass Spectrometry of the Diversity of Endophytes and Rhizobacteria Cultured from the Maize Microbiome

Michael G. LaMontagne¹, Phi L. Tran¹, Alexander Benavidez², Lisa D. Morano²

¹ Department of Biology and Biotechnology, University of Houston – Clear Lake, Houston, TX, United States

² Natural Sciences, University of Houston – Downtown, Houston, TX

Corresponding Author:

Michael Geary LaMontagne¹

2700 Bay Area Blvd., Houston, TX, 77058, United States

Email address: lamontagne@uhcl.edu

Abstract

Many endophytes and rhizobacteria associated with plants support the growth and health of their hosts. The vast majority of these potentially beneficial bacteria have yet to be cultured, in part because of the cost of identifying bacterial isolates. Matrix-assisted laser desorption – time of flight (MALDI-TOF) has enabled high throughput “culturomics” studies of host-associated microbiomes but analysis of mass spectra generated from plant-associated bacteria requires optimization. In this study, we aligned mass spectra generated from endophytes and rhizobacteria isolated from heritage and sweet varieties of *Zea mays*. Multiple iterations of alignment attempts identified a set of parameters that sorted 114 isolates into 60 coherent MALDI-TOF taxonomic units (MTUs). These MTUs corresponded to strains with practically identical (> 99%) 16S rRNA gene sequences. Mass spectra were used to successfully train a machine learning algorithm that perfectly classified isolates. The 60 MTUs provided > 70 % coverage of aerobic, heterotrophic bacteria readily cultured with nutrient rich media from the maize microbiome and allowed prediction of the total diversity recoverable with that particular cultivation method. *Acidovorax* sp., *Pseudomonas* sp. and *Cellulosimicrobium* sp. dominated the library generated from the rhizoplane and the heritage variety contained a relatively high number of MTUs specific to that niche. This suggests a rapid, inexpensive method of describing the types of bacteria cultured from niches within maize microbiomes and the coverage achieved with a cultivation strategy.

Introduction

Current agricultural practices will not meet the nutritional needs of a population that will reach nine billion people by the middle of this century (Anand et al. 2013). In parallel, climate change will increase extreme weather events, including drought (Dai 2011; Trenberth et al. 2014), and urbanization will eliminate arable land (Song et al. 2015). Microbial products can mitigate this food crisis by increasing crop yield (Tkacz & Poole 2015) and helping crops tolerate drought and marginal soils (Brígido & Glick 2015). The plant microbiome abounds with endophytes and plant growth promoting rhizobacteria (PGPR) that can help plants recover more nutrients from the soil and tolerate stressors like drought (Barnawal et al. 2013; Bresson et al. 2014). PGPR can also control plant pathogens (Chowdhury et al. 2013), promote beneficial mycorrhizae colonization (Labbé et al. 2014) and produce potentially valuable secondary metabolites (Brader et al. 2014; Kumar et al. 2012; Raaijmakers & Mazzola 2012); however, the vast majority of microbes associated with agrosystems, have yet-to-be cultured. This limits the development of microbial products from PGPR and manipulative experiments with defined and representative plant microbiomes that can facilitate hypothesis testing, as demonstrated with maize (Niu et al. 2017).

Given its long history of artificial selection and its importance as a global food crop, the maize microbiome is of particular interest. Maize stems (Kämpfer et al. 2016a; Kämpfer et al. 2016b) have yielded novel bacterial species and maize kernels host diverse bacterial endophytes (Rijavec et al. 2007). A few studies have compared the diversity of libraries of bacteria isolated from different varieties of different niches within the maize microbiome (McInroy & Kloepper 1995). Further, although Bt and non-Bt modern maize showed no differences in their bacterial communities (Mashiane et al. 2017), maize varieties may support different microbiomes. For example, a comparison of libraries generated from a wild maize and a more modern variety yielded three PGPR species from the wild maize that were not recovered from the modern variety (Mousa et al. 2015) and ancient maize yielded a bacterial endophyte that inhibits multiple fungal pathogens (Shehata et al. 2016). A deeper understanding of interactions between maize varieties and the microbiomes they host in their endosphere and rhizosphere could inform searches for PGPR.

Matrix-assisted laser desorption – time of flight (MALDI-TOF) appears well suited for analysis of libraries isolated from plant microbiomes (Ghyselinck et al. 2013). MALDI-TOF systems provide strain-level identification of microbes (Ahmad et al. 2012; Sauer et al. 2008; Singhal et al. 2015) for pennies an isolate. These systems use pattern matching between mass and reference spectra; however, these databases have poor representation of environmental microbes. This paucity of representation limits MALDI-TOF effectiveness (Singhal et al. 2015). Compilations of spectra are available (Böhme et al. 2012; Murugaiyan et al. 2018; Rau et al. 2016). These compilations are typically species-specific databases of mass spectra, including databases for *Mycobacterium kansasii* (Murugaiyan et al. 2018) and *Vibrio* species (Erler et al. 2015) or for particular systems, like spacecraft (Seuylemezian et al. 2018). Further, custom databases require proprietary software and do not facilitate investigators sharing mass spectra, which would benefit science. To address this, several teams have developed applications for microbial identification by MALD-TOF by matching of spectra to peaks inferred from genomic (Tomachewski et al. 2018) or proteomic databases (Cheng et al. 2018) and cluster analysis allows for rapid

dereplication without a reference library (Clark et al. 2018; Dumolin et al. 2019). Web applications for this analysis are available (LaMontagne et al. 2017); however, data analysis of mass spectra generated from bacteria isolated from the plant microbiome needs improvement (Huschek & Witzel 2019).

In this study, we generated a library of readily-culturable bacteria from the endosphere and rhizosphere of two varieties of maize (*Zea mays*): an agronomically important hybrid and a heritage variety. Isolates were sorted, using custom scripts, into MTUs based on similarity of mass spectra generated by MALDI-TOF. These clusters corresponded to strains with nearly identical rRNA gene sequences, species identified with a commercial mass spectra database, and provided > 70 % coverage of the aerobic, readily-culturable bacteria of the maize microbiomes we sampled. In terms of these clusters, the rhizoplane of the heritage variety yielded the more MTUs that were unique to that particular niche than libraries generated from the endosphere and sweet variety.

Materials & Methods

Sampling Maize Microbiome

Two maize varieties were chosen for this experiment, a ‘Heritage’ and a ‘Sweet’ variety. The Heritage variety was Dent Earth Tones Corn (Botanical Interest Inc., Broomfield, CO). This is a heritage variety selected for propagation based on its colorful kernels. The Sweet variety, Bodacious Hybrid Sweet Corn (American Seed Co, Spring Grove, PA), has been highly bred for both sweetness and disease resistance. This modern variety has the designation “se” which stands for sugar enhancer and “R/M” indicating disease resistance to multiple maize diseases.

Plants were grown inside at the University of Houston-Downtown in fall of 2017. Two seeds per variety were planted in 3.5 inch x 3.5 inch pots containing LadyBug Vortex Potting Soil® (New Earth, Conroe, TX). Plants were grown at room temperature in a two-tier plant growing unit with grow lights. The plants for this experiment were all on the same tier in two batches. Grow lights were set approximately 0.7 meters above the plants and set for 12/12 hour light/cycle. Pots were watered every 2-3 days. One batch, designated “New” was harvested 60 days after sowing. Another batch, designated “Old” was harvested 120 days after sowing. All plants were brought to the University of Houston - Clear Lake for plant measurements and microbial extraction.

Investigation of both maize varieties began with the cutting of plants into shoots and root. Soil was knocked from the roots and the fresh weight of shoots and roots was recorded. The height of each shoot was recorded from the soil line to the longest shoot tip. Roots were transferred into a 50 ml conical tube and wetted with 40 ml of 5 % DMSO prepared in sterile distilled water. The tubes were then vortexed for 1 min and centrifuged at 4,800 x g, in a swing-out rotor, for 10 min at room temperature. The roots were recovered with tweezers and blotted dry to obtain wet weights. The supernatant was discarded and the remaining slurry (~ 5 ml) was transferred to microtubes with a wide orifice pipet.

Endophytes were recovered from stem sections that were cut into lengths that weighed between 0.20 and 0.25 grams. These pieces were submersed in 70% ethanol for 60 seconds followed by flaming to remove ethanol. The pieces were then rinsed in sterile distilled water for 60 seconds. The pieces were then cut into approximately 1mm fragments and ground to a slurry in 2 ml of sterile water, using a surface sterilized mortar and pestle (Johnston-Monje et al. 2014).

Slurries generated from roots and shoots were serially diluted in sterile phosphate buffered saline and inoculated on half strength tryptic soy broth (EMD Millipore, Billerica MA) solidified with 15 g per liter with agar (AmericanBio, Natick, MA) to make TSA plates. The track dilution method was used to spread dilutions (BD et al. 1997) and plates were allowed to dry before being sealed with parafilm and incubated at 30°C for 48 hours. Colonies were picked from dilutions (generally the 10⁻⁴ dilution) that yielded isolated colonies. To maximize diversity, morphology was considered in picking colonies. Representative morphologies were selected from each track. Isolates were re-streaked to fresh plates twice for purity.

Microbial Identification - MALDI-TOF

Isolates were prepared for MALDI-TOF analysis with the ethanol inactivation and formic acid extraction protocols recommended by Bruker Scientific (Billerica, MA). This approach follows the recommendations of Freiwald and Sauer (Freiwald & Sauer 2009), except 50 µl volumes of formic acid and acetonitrile were used. Briefly, bacteria were cultured overnight at 30 °C on TSA plates. Resulting colonies were then suspended in 300 µl of HPLC-grade water and inactivated by adding 900 µl HPLC-grade ethanol. The resulting slurry was stored at 4 °C for one week. The ethanol inactivated cells were then recovered by centrifugation (16,000 x g, 2 min) and the resulting pellet was extracted with formic acid/acetonitrile to yield a formic acid extract.

Matrix solution, 10 mg/mL a-cyano-4-hydroxycinnamic acid (HCCA, Sigma, St. Louis, MO) in 50% acetonitrile, 47.5% water, and 2.5% trifluoroacetic acid (ACN, Sigma) was prepared fresh for each use. Bacterial test standard (BTS, Bruker p/n 8255343) was dissolved in 50 µL of 50% aq. ACN, 2.5% TFA following the manufacturer instructions. MALDI target plate (MSP 96 polished steel target, Bruker p/n 8280800) was washed the trifluoroacetic acid and 70% ethanol as recommended by Bruker. On one or two spots on the target, 1 µl of formic acid extract was spotted on the target followed by 1 µl of matrix solution. Two BTS spots were prepared by applying 1 µl of BTS solution followed by 1 µl of matrix solution to each spot. Two matrix blank spots were included on each target as negative controls. Targets were allowed to dry at room temperature for approximately 15 minutes and they were shipped overnight, with an ice pack, to the Proteomics and Mass Spectrometry Core Facility at the Huck Institute (The Pennsylvania State University, University Park, PA 16802). Positive-ion mass spectra were acquired on a Bruker Ultraflexextreme MALDI TOF/TOF mass spectrometer. Linear detection mode with the following parameters was used: pulsed ion extraction 170 ns; Ion source 1 25 kV; Ion source 2 94% of Ion source 1, and Lens 32% of Ion source 1. Matrix suppression (deflection) was set to 1500 m/z. The laser repetition rate was 667 Hz; Smartbeam parameter set to “3_medium”. Real-time smoothing was Off, baseline offset 0.2%, analog offset 2.1 mV, and the detection was set for low mass range, 1,880 – 20,000 Da. The target was moved in a random

walk, complete sample pattern; 50 shots were fired at each raster spot limited to a 2000-mm diameter.

Mass calibration was performed with BTS as the standard; quadratic calibration curve was constructed based on m/z values of 8 calibrants over 3637 to 16952 m/z range. Mass spectra were smoothed using 10 cycles of SavitzkyGolay, with width 2 m/z; and baseline-subtracted using TopHat algorithm. Mass List Find parameters were as follows: peak detection algorithm Centroid, S/N threshold 2, minimum intensity threshold 600, max number of peaks 300, peak width 4 m/z, height 90%. The processed mass spectra were loaded into the MALDI Biotyper Version 3.1 (build 66) software (Bruker) and searched against a Bruker Taxonomy library containing 7,854 entries.

Mass spectra were also analyzed by cluster analysis using an R script that implemented functions in several packages (MALDIquant (Gibb & Strimmer 2012), PVLust(Suzuki & Shimodaira 2006)...). This script included two optimization loops that iteratively sampled random values, within specified ranges, for seven parameters: half-window for smoothing, baseline removal, half-window for alignment, tolerance of alignment, signal to noise ratio (SNR) for alignment, half-window for peak detection and SNR for peak detection. The first loop identified the parameters that optimized the number of peaks shared, as Jaccard coefficients calculated with the philentropy (Drost 2018) between pairs of average mass spectra generated from BTS. Output from the first optimization was passed to a step then conducted quality control analysis to identify noisy spectra. On the spectra that passed quality control, the second loop selected the parameters that minimized the overlap in cosine similarity values, calculated following (Strejcek et al. 2018), between closely related and distantly related isolates. These cosine similarities were weighted by Jaccard coefficients normalized after with the formula $y = y_0 + x/(x + 0.2)$, where x is Jaccard coefficient, y_0 is the average cosine similarity when $x = 0$ and y is a predicted cosine value. The script then performed cluster analysis to define MALDI-TOF taxonomic units (MTUs), rarefaction analysis using the iNext package (Hsieh et al. 2016) and trained a machine learning algorithm using the RWeKa package (Hornik et al. 2009). This script is presented in one R markdown file (Supplemental Data). Raw mass spectra and Bruker system identifications are available as dataset MSV000086274 in MassIVE (Deutsch et al. 2016).

Curve fitting and ANOVA testing of significance of correlations were completed with SigmaPlot for Windows version 13.0 (Systat Software, San Jose, CA).

Microbial Identification - 16S rRNA gene sequencing

Representative isolates of MTUs that were not identified at the species level using the Biotyper database and software (Biotyper score < 2), were identified by 16S rRNA gene sequencing. DNA was extracted from isolates cultured overnight in tryptic soy broth (30 °C, 200 rpm) with the Puregene kit, following the manufacturer's protocol (Qiagen, Carlsbad CA). A near intact fragment of the 16S rRNA gene was amplified with ReadyMade™ Primers 16SrRNA For and 16srRNA Rev supplied by IDT (Coralville, IA) using DreamTaq Hot Start DNA polymerase, following the manufacturer's protocol (ThermoFisher, Waltham, MA). PCR reactions were preheated (94 °C, 2 m) and then cycled 32 times through the following steps: denaturing (94 °C, 30 s), annealing (56 °C, 30 s) and extension (72 °C, 90 s). The final extension step was extended

for 5 minutes and product size was confirmed by electrophoresis with a FlashGel™ DNA Kit (Lonza, Basel, Switzerland). PCR products were then purified with the DNA Clean & Concentrator kit, following the manufacturer's protocol (Zymo, Irvine, CA). Purified fragments were sequenced with the above ReadyMade™ Primers, using Sanger technology, by Lone Star Laboratory (Houston, TX).

Single-pass, 16S rRNA sequences were manually curated in BioEdit(Hall 1999), checked for chimeras with Decipher (Wright et al. 2012) and aligned against curated sequences with SINA (Pruesse et al. 2012). ModelFinder (Kalyaanamoorthy et al. 2017) was used within the IQ-TREE environment (Trifinopoulos et al. 2016) to select the appropriate phylogenetic maximum likelihood model. Bootstrap values were calculated with UFBoot2 (Hoang et al. 2017) and a tree was generated with Newick Display 1.6 (Junier & Zdobnov 2010). Decipher, ModelFinder, UFBoot2 and Newick Display were all run with default parameters. The 16S sequences were uploaded to GenBank and assigned accession numbers MW092913 - MW092913.

Results

Iteratively trying different parameters, like tolerance and SNR for aligning mass spectra and detecting peaks respectively, identified a set that optimized the number of peaks shared between spectra generated from BTS references spotted on separate targets. The number of peaks shared between reference spectra reached an asymptote after a dozen iterations (Fig. S1). Extending the number of iterations of the optimization loop to 2,000 identified parameters that corresponded to the highest Jaccard similarity coefficients (0.952), in terms of peaks matched between spectra generated from BTS. Pairwise comparisons of these BTS spectra shared 20/21 peaks and showed cosine similarities that ranged from 0.963 – 0.970. Jaccard similarity coefficients between these reference spectra showed a modal relationship with the number of peaks detected (Fig. S2 top). The zenith of this relationship occurred between 389 – 406 total peaks detected from the 174 spectra that passed initial quality control within the Bruker system. The parameters that that yielded the most reproducible peaks were selected for identification of quality spectra, as defined by the number of the SNR for the ten largest peaks and the number of peaks detected. Quality control analysis, using a threshold of a median SNR of 15 for the 10 largest peaks and detection of at least 13 peaks, pruned mass spectra generated from 8 of 122 isolates from subsequent cluster analysis. The second optimization loop found, for pairs of isolates that had practically identical 16S rRNA sequences (> 99 %), the average Jaccard coefficient for comparisons reached a zenith at 498 peaks (Fig S2 bottom).

Cosine similarities calculated with the parameters that showed the greatest discrimination between comparisons within species than between species, suggested a hyperbolic relationship with the Jaccard similarities (Fig. 1). Most (6472/6670) of the pairwise comparisons showed Jaccard similarity coefficients of less than 0.2. In that range, cosine similarities averaged ($\pm$ SD) 0.27 ± 0.10 . For the 192 pairwise comparisons that showed Jaccard similarities greater than or equal to 0.2, cosine similarities averaged 0.81 ± 0.16 . The hyperbolic model suggested that Jaccard similarities of 0.2 corresponded to cosine similarities of about 0.69.

The percent identities of pairwise comparisons 16S rRNA genes fit a sigmoidal model with weighted cosine similarities (Fig. 2). The fit had a modest correlation coefficient ($r^2= 0.445$) and

analysis of variance suggested the trend was highly probable ($P < 0.0001$); however, most (259/325) of the pairwise comparisons showed percent identities of 16S rRNA genes of less than 90%. The midpoint of this sigmoidal model corresponded to a cosine similarity of 0.66. Cosine similarities of > 0.66 consistently corresponded to a percent identity of ribosomal sequences ranging from 99 – 100 % (Fig. 2).

Inspection of Figure 2 suggested that cosine similarity of 0.65 would distinguish species. Using that threshold, clustering of a library of 114 isolates into MALDI-TOF taxonomic units (MTUs) defined 60 MTUs. Of these, 25 clusters contained more than one isolate (replicated MTUs) and 35 were singletons. Extrapolation from rarefaction analysis predicted that a library of 228 isolates would contain 84 MTUs, with a confidence interval of 70 – 98 (Fig. 3 top). This library of isolates corresponds to 70 % coverage, with a confidence interval of 62 – 77 % (Fig. 3 bottom). Further extrapolation predicted that a library of 228 isolates would provide 78 – 95 % coverage.

Partial sequencing of 16S rRNA genes yielded sequences that were used to classify isolates into 18 phylotypes that corresponded to different genera (Fig. 4). These phylotypes generally appeared robust, as assessed by bootstrap values $> 90\%$. The exception was separation of the branch containing isolates that classified as *Paenibacillus* and *Bacillus* species, which bootstrapping supported only 76% of iterations. Within this Bacillales order, and in terms of rRNA gene sequences, isolate H1OZ89 showed high similarity to a phosphate solubilizing *Paenibacillus* species previously isolated from rape (Ding et al. 2019). Also, within this order, isolate H3OZ122 showed high similarity to a microplastics degrading *Bacillus* strain recently isolated from the Yellow Sea (Wang et al. 2019), isolate S4OZ125 showed high similarity to *Bacillus aciditolerans*, a recently described novel species isolated from a rice field (Ding et al. 2019) and H3OZ107 showed similarity to a *Staphylococcus hominis* strain isolated from a blood sample (Fig. 4). These isolates (H1OZ89, H3OZ122, S4OZ125 and H3OZ107) clustered into separate MTUs, which suggests the similarity threshold defined MTUs corresponds to species that occupy different niches.

Consistent with MTUs that classified within the Bacillales order, the topology of Figure 4 generally showed congruence with MTUs. For example, MTUs 33, 34, 46 and 51 each contained two isolates that both clustered together in the same leaf on the tree. For the seven isolates of the *Acidovorax* genera shown in Figure 4, the threshold for clustering appeared conservative. The rRNA sequences generated from seven isolates showed $> 99\%$ identity with species of *Acidovorax*, and each other, but were separated into three MTUs (29, 35 and 54). MTU 35 corresponded to a branch in Figure 4 supported with 94 % bootstrapping value, which suggests that cluster is coherent. On the other hand, MTUs 25 and 54 appear to have isolates of the same species, split into two clusters.

Only 36 of the 114 isolates (32 %) were identified at the species level with probable confidence, as defined by a classification score of > 2 with the Biotyper system, and 66/114 (58%) were identified at the genera level. Matching to reference spectra did not appear to depend on the quality of the spectra, as classification scores showed no relationship with the SNR of spectra. The set of identified species contained 35 isolates that classified as bacteria and one isolate that classified as a fungal species (*Trichosporon mucoides*). Of the 35 bacterial isolates 6 were

singletons. The 29 replicated isolates were classified into 11 species. These isolates generally corresponded to specific MTUs. For example, the 6 isolates reliably identified as *Pseudomonas corrugata* all clustered into MTU 20 (Fig. 5). Similarly, isolates identified by the Bruker system as *Pseudomonas mendocina*, *Enterobacter cloacae*, *Serratia marcescens*, *Bacillus firmus*, *Cellulosimicrobium cellulans*, *Pseudomonas citronellolis*, *Rhizobium radiobacter* and *Acidovorax facilis* clustered within MTUs 34, 33, 3, 21, 8, 4, 25 and 29 respectively (Fig 5). Bootstrap values, for iterations of cluster analysis, generally supported the congruence of these MTUs with species identifications. Bootstrap values for these clusters exceeded 95 %, with the exception of MTU 34 (Fig. 5). However, the dendrogram was not coherent with respect to the phylogeny of the isolates. In particular, *Pseudomonas* species were dispersed throughout the tree and appeared within a branch that contained Gram positive species (Fig. 5) and, for the three isolates the system identified as *P. putida*, the threshold used to define MTUs appeared conservative. These isolates were split into two MTUs (Fig. 5). The lone isolate in MTU 26 showed little similarity, in terms of cosine similarity, to any other isolate.

Machine learning resolved the apparent discrepancy for the isolates that clustered into different MTUs but were identified as one species by the Bruker system. To train the machine learning algorithm, MTUs were identified by the Bruker system and 16S sequencing. When the Bruker and 16S identification differed, the 16S identification was used but this was rare. Identification by the Bruker system and 16S sequencing agreed at the genera level, for Bruker scores greater than or equal to 1.6; however, the two approaches showed some discordance at the species level. For example, MTU 33 was reliably identified (Bruker score = 2.5) as *Enterobacter cloacae* but these isolates clustered with *Enterobacter ludwigii* in Figure 4. MTU 34 was probably identified (Bruker > 2) as *Pseudomonas mendocina* but these isolates clustered with *P. alcaliphila* in Figure 4. MTU 35 were identified by the Bruker system as *Acidovorax facilis* with scores > 2.2 but a representative of that cluster showed high similarity, in terms of 16S sequence, to *A. wautersii* (Fig. 4). This MTU was defined as *A. wautersii*. When species were split into two MTUs, they were pooled. For example, the three isolates identified, at the probable level, as *P. putida* by the Bruker system were defined as that species, even though they belonged to two MTUs (Fig. 5). MTUs 54 and 29 were identified as one clade *Acidovorax sp.* based on Figure 4. After discarding 34 singletons, this classification scheme provided at least genera level identities for all but two of the remaining isolates. The machine learning algorithm classified 100% of the 80 replicated isolates, and the two BTS references, correctly as 22 MTUs. This corresponds to a perfect kappa score.

A combination 16S sequencing and MALDI TOF methods identified distinct MTUs that were relatively abundant to particular niches or plant varieties. Libraries generated from the rhizoplane contained the highest number of replicated MTUs overall (Fig. 6). Of those 20 MTUs, 14 did not appear in libraries generated from the endosphere. These rhizoplane-specific MTUs included numerous representatives of two *Acidovorax sp.* and *Flavobacterium sp. Niveispirillum sp.*, *Pseudoxanthomonas sp.*, *Rhizobium radiobacter* and *Shinella sp.* (Fig. 6). MTUs specific to the endosphere included a yet-to-be identified cluster (MTU02), *Serratia sp.*, *Arthrobacter sp.*, and *Pseudomonas sp.* (Fig. 6). MTUs identified as *Pseudomonas sp.*, *Cellulosimicrobium sp.*, *Bacillus sp.*, and *Enterobacter sp.* were found in both the rhizoplane and the endosphere (Fig. 6). Libraries generated from the heritage variety contained more MTUs (11) unique to that variety than libraries generated from the sweet variety (Fig. S3).

Discussion

Optimization of cluster analysis of mass spectra generated from a library of bacteria isolated from maize microbiomes produced coherent MTUs that corresponded to species defined by 16S rRNA gene sequencing, MALDI Biotyper database and niches typically occupied by species related to these MTUs. The library of readily culturable bacteria isolated from the rhizoplane of the heritage variety contained the most MTUs specific to a particular niche. That is strains that abound in libraries generated from that sample and were not common in libraries generated from the endosphere of that variety or from rhizoplane and endosphere of the sweet variety. The rhizoplane and endosphere have many species in common, but it is logical that the rhizoplane community is more diverse given its direct contact with soil, which are diverse systems (Howe et al. 2014). In fact, endophyte communities may be a subset of the more extensive rhizosphere communities (Long et al. 2010).

The relatively high abundance of *Acidovorax* and *Flavobacteria* in libraries generated from the rhizoplane (Fig. 6) is consistent with previous culture-dependent surveys of plant microbiomes. *Acidovorax* is commonly found in both soil and plants (Long et al. 2010) and *Flavobacteria* genera includes a terrestrial clade associated with plant roots plant roots (Kolton et al. 2013). However, these genera are not consistently associated with maize roots, as assessed by metagenomic analysis (Niu et al. 2017). We also isolated *Rhizobium radiobacter* (formerly *Agrobacterium tumefaciens*). This soil bacteria interacts with plants and strains of this species show potential for bioremediation applications (Deepika et al. 2016) and novel species of this genera have been isolated from maize roots (Gao et al. 2017). *Microbacterium* has been isolated specifically from maize kernels (Rijavec et al. 2007). This endophyte may have been in the seed when it was sown.

The library generated from the maize endosphere included species that colonize plants. The most abundant MTU specific to the endosphere (MTU02) was not identifiable with the Bruker system and not archived for future analysis. Other MTUs that abounded in libraries generated from the endosphere included *Serratia*, *Arthrobacter* and *Pseudomonas* species. *Serratia marcescens* is a common endophyte in cotton roots (McInroy & Kloepper 1995) and rice (Tan et al. 2001). Several *Serratia* species associate with maize roots (Mosimann et al. 2017) and *S. marcescens* can cause maize whorl rot (Wang et al. 2015). *Arthrobacter spp.* are readily isolated from cotton stems (McInroy & Kloepper 1995) and identified as an endophyte in maize sap (Ali et al. 2018).

MTUs found in both the libraries generated from the rhizoplane and endosphere included genera frequently associated with plant hosts. *Pseudomonas sp.* are readily cultured from the rhizoplane of corn on selective media (Mosimann et al. 2017). *Cellulosimicrobium sp.* can stimulate plant growth and act as a biocontrol for barley (Nabti et al. 2014). Recently the genome of a *Cellulosimicrobium sp.* strain isolated from endosphere of a perennial grass was sequenced because of its growth promoting properties (Eida et al. 2020). *Bacillus sp.* are ubiquitous in soils and widely used as microbial treatments of seeds (Rocha et al. 2019) and *Enterobacter sp.* can promote the growth of maize (Naveed et al. 2014) and appear a core member of that host's microbiome (Niu et al. 2017).

Many of the potentially beneficial rhizobacteria and endophytes presented above appear associated with only the heritage and not the sweet variety. For example, *Cellulosimicrobium* and *Serratia* sp. was common to only the heritage variety and *Enterobacter* and *Flavobacteria* sp. were common to only the sweet variety. This can inform microbial discovery efforts, where the goal is to isolate strains of a particular genera; however, for development of microbial products to improve the productivity and sustainability of maize, the variety cultivated by growers should be considered. Plant varieties form close associations with particular beneficial rhizobacterial strains (Batstone et al. 2020).

Identification of maize-associated bacteria involved leveraging species identification the Bruker system with 16S rRNA gene sequencing. This hybrid approach is widely used in applications of MALDI-TOF to environmental microbiology, where many of the isolates will not match the commercial database. For this study, the percentage of bacteria identified to the probable species level (32%) with the Biotyper system was lower than previous reports (73%) for soil bacteria (Strejcek et al. 2018) but in the range reported for a library generated from the rhizosphere of horseradish (Uhlik et al. 2011) and bacteria isolated from seawater (Timperio et al. 2017). We also observed discrepancies between species identifications for the two approaches, as reported previously and confirmed recently (Pandey et al. 2019). This could reflect the paucity of environmental isolates in the reference library we queried. The confidence of identification shows a hyperbolic relationship with the number of reference spectra in the database and approaches an asymptote at about 50 main spectra for each strain (Erler et al. 2015). The commercial database we queried only contains a few main spectra for most species.

With cluster analysis, we can begin to understand the diversity of rhizobacteria cultured with any particular technique from a plant's microbiome. Clustering of mass spectra, following optimization of alignment, identified MTUs that corresponded to highly similar (> 99%) 16S rRNA gene sequences and species defined by a commercial database. The cosine similarity (0.65) selected as a threshold for defining MTUs was lower than threshold (0.79) reported by Strejcek et al. (Strejcek et al. 2018) but the coherence of the clusters, in terms of the consistency with clusters by 16S rRNA sequencing defined by a commercial spectra database and, supports this lower threshold. A more conservative threshold would split clusters that cannot be resolved with 16S rRNA sequencing, which is the current standard for microbial identification (Edgar 2018). High bootstrap support for a dendrogram generated from mass spectra (Fig. 5) suggests the data analysis pipeline used to generated these clusters is robust. This suggests an approach comprehensively compare the types of bacteria in libraries generated from different parts of the maize microbiome and test hypotheses about the efficacy of advanced culturing techniques, such as *in situ* cultivation (Berdy et al. 2017), improved media formulations (Tanaka et al. 2014) and gnotobiotic systems (LaMontagne 2020). Machine learning also supported the clusters defined herein, and this could lead to an alternative method of identification (De Bruyne et al. 2011). Indeed, dozens of machine learning algorithms have recently been applied to this task (Weis et al. 2020) and our analysis suggests Rweka shows promise; however, development of such an identification system would require a challenge dataset generated externally, preferably on a different instrument (Clark et al. 2020).

The reliance on 16S sequences limits the scope of this study. Phylogenetic analysis of the 16S rRNA gene benefits from a vast public database and user-friendly data analysis packages;

however, analysis of this one gene lacks resolution. For example, diverse *Bacillus* species share identical rRNA gene sequences. Multilocus sequence typing (MLST) provides a higher resolution approach (Maiden et al. 1998) but MLST can fail to accurately estimate phylogenetic relationships between bacteria (Tsang et al. 2017). Whole genome sequencing provides strain-level identification (Salipante et al. 2015) but costs hundreds of dollars per isolate. Continued decreasing costs per bp of next generation sequencing (Park & Kim 2016) and application of low-cost, third generation sequencers (Petersen et al. 2019) suggests whole genome sequencing will soon be standard for clinical applications (Anis et al. 2018); however, the true cost of whole genome includes library preparation, skilled labor and computational costs (Sboner et al. 2011). These costs limit the application of this approach to high-throughput microbial discovery programs.

The pooling of isolates collected at different stages of growth and isolated on different media into one library limits the scope of this study and precludes testing of hypothesis about the controls of the diversity of bacteria associated with maize. Defining the diversity of maize microbiome, within different locations of the host and between different varieties of maize, would require generation of much larger libraries than is feasible with culture-dependent techniques. Metagenomic analysis is better suited testing hypotheses and provides a more complete census of the microbial communities in nature (Handelsman 2004); however, this dataset revealed trends consistent with metagenomic studies, ecological theory and published observations.

Conclusions

MALDI-TOF facilitated the classification of bacteria isolated from the maize microbiome into 60 MTUs. This suggests a rapid, inexpensive method of describing the types of bacteria cultured from plant microbiomes with freely available spectra alignment and machine learning packages.

Acknowledgements

The authors thank Dr. Tatiana Laremore at the Proteomics and Mass Spectrometry Core Facility (Pennsylvania State University) for MALDI-TOF mass spectra acquisition, MBT microorganism identification, and help with the manuscript preparation. The authors thank Torri Fugate-Mullins (University of Houston – Clear Lake) for help with microbial isolation and manuscript preparation.

References

- Ahmad F, Babalola O, and Tak H. 2012. Potential of MALDI-TOF mass spectrometry as a rapid detection technique in plant pathology: identification of plant-associated microorganisms. *Analytical and Bioanalytical Chemistry* 404:1247-1255. 10.1007/s00216-012-6091-7
- Ali S, Isaacson J, Kroner Y, Saldias S, Kandasamy S, and Lazarovits G. 2018. Corn sap bacterial endophytes and their potential in plant growth-promotion. *Environmental Sustainability* 1:341-355. 10.1007/s42398-018-00030-4
- Anand R, Grayston S, and Chanway C. 2013. N₂-fixation and seedling growth promotion of Lodgepole Pine by endophytic *Paenibacillus polymyxa*. *Microbial Ecology* 66:369-374. 10.1007/s00248-013-0196-1
- Anis E, Hawkins IK, Ilha MRS, Woldemeskel MW, Saliki JT, and Wilkes RP. 2018. Evaluation of targeted next-generation sequencing for detection of bovine pathogens in clinical samples. *Journal of Clinical Microbiology* 56:00399-00318. 10.1128/jcm.00399-18
- Barnawal D, Maji D, Bharti N, Chanotiya C, and Kalra A. 2013. ACC deaminase-containing *Bacillus subtilis* reduces stress ethylene-induced damage and improves mycorrhizal colonization and rhizobial nodulation in *Trigonella foenum-graecum* under drought stress. *Journal of Plant Growth Regulation* 32:809-822. 10.1007/s00344-013-9347-3
- Batstone RT, O'Brien AM, Harrison TL, and Frederickson ME. 2020. Experimental evolution makes microbes more cooperative with their local host genotype. *Science* 370:476-478. 10.1126/science.abb7222
- BD J, KL H, MM H, and MS G. 1997. Simplified agar plate method for quantifying viable bacteria. *BioTechniques* 23:648-650.
- Berdy B, Spoering AL, Ling LL, and Epstein SS. 2017. *In situ* cultivation of previously uncultivable microorganisms using the ichip. *Nature Protocols* 12:2232. 10.1038/nprot.2017.074
- Böhme K, Fernández-No IC, Barros-Velázquez J, Gallardo JM, Cañas B, and Calo-Mata P. 2012. SpectraBank: An open access tool for rapid microbial identification by MALDI-TOF MS fingerprinting. *Electrophoresis* 33:2138-2142. 10.1002/elps.201200074
- Brader G, Compant S, Mitter B, Trognitz F, and Sessitsch A. 2014. Metabolic potential of endophytic bacteria. *Current Opinion in Biotechnology* 27:30-37. 10.1016/j.copbio.2013.09.012
- Bresson J, Vasseur F, Dauzat M, Labadie M, Varoquaux F, Touraine B, and Vile D. 2014. Interact to survive: *Phyllobacterium brassicacearum* improves Arabidopsis tolerance to severe water deficit and growth recovery. *PLoS ONE* 9:107607. 10.1371/journal.pone.0107607
- Brígido C, and Glick B. 2015. Phytoremediation using rhizobia. In: Ansari AA, Gill SS, Gill R, Lanza GR, and Newman L, eds. *Phytoremediation*: Springer International Publishing, 95-114.
- Cheng D, Qiao L, and Horvatovich P. 2018. Toward spectral library-free matrix-assisted laser desorption/ionization time-of-flight mass spectrometry bacterial identification. *Journal of Proteome Research* 17:2124-2130. 10.1021/acs.jproteome.8b00065

- Chowdhury SP, Dietel K, Rändler M, Schmid M, Junge H, Borriess R, Hartmann A, and Grosch R. 2013. Effects of *Bacillus amyloliquefaciens* FZB42 on lettuce growth and health under pathogen pressure and its impact on the rhizosphere bacterial community. *PLoS ONE* 8:68818. 10.1371/journal.pone.0068818
- Clark CM, Costa MS, Sanchez LM, and Murphy BT. 2018. Coupling MALDI-TOF mass spectrometry protein and specialized metabolite analyses to rapidly discriminate bacterial function. *Proceedings of the National Academy of Sciences* 115:4981-4986. 10.1073/pnas.1801247115
- Clark CM, Murphy BT, and Sanchez LM. 2020. A call to action: the need for standardization in developing open-source mass spectrometry-based methods for microbial subspecies discrimination. *mSystems* 5:00813-00819. 10.1128/mSystems.00813-19
- Dai A. 2011. Drought under global warming: a review. *Wiley Interdisciplinary Reviews: Climate Change* 2:45-65. 10.1002/wcc.81
- De Bruyne K, Slabbinck B, Waegeman W, Vauterin P, De Baets B, and Vandamme P. 2011. Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning. *Systematic and Applied Microbiology* 34:20-29. 10.1016/j.syapm.2010.11.003
- Deepika KV, Raghuram M, Kariali E, and Bramhachari PV. 2016. Biological responses of symbiotic *Rhizobium radiobacter* strain VBCK1062 to the arsenic contaminated rhizosphere soils of mung bean. *Ecotoxicology and Environmental Safety* 134:1-10. 10.1016/j.ecoenv.2016.08.008
- Deutsch EW, Csordas A, Sun Z, Jarnuczak A, Perez-Riverol Y, Ternent T, Campbell DS, Bernal-Llinares M, Okuda S, Kawano S, Moritz RL, Carver JJ, Wang M, Ishihama Y, Bandeira N, Hermjakob H, and Vizcaino JA. 2016. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Research* 45:D1100-D1106. 10.1093/nar/gkw936
- Ding MJ, Shang NJ, Xiao ZX, Shao F, Liu L, Huang Y, Zhou LK, Zhou JH, and Zhang Y. 2019. *Bacillus aciditolerans* sp. nov., isolated from paddy soil. *International Journal of Systematic and Evolutionary Microbiology* 69:1155-1161. 10.1099/ijsem.0.003285
- Drost H-G. 2018. Philentropy: information theory and distance quantification with R. *Journal of Open Source Software* 3:765.
- Dumolin C, Aerts M, Verheyde B, Schellaert S, Vandamme T, Van der Jeugt F, De Canck E, Cnockaert M, Wieme AD, Cleenwerck I, Peiren J, Dawyndt P, Vandamme P, and Carlier A. 2019. Introducing SPeDE: high-throughput dereplication and accurate determination of microbial diversity from matrix-assisted laser desorption–ionization time of flight mass spectrometry data. *mSystems* 4:00437-00419. 10.1128/mSystems.00437-19
- Edgar RC. 2018. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* 34:2371-2375. 10.1093/bioinformatics/bty113
- Eida AA, Bougouffa S, Alam I, Saad MM, and Hirt H. 2020. Complete genome sequence of the endophytic bacterium *Cellulosimicrobium* sp. JZ28 isolated from the root endosphere of the perennial desert tussock grass *Panicum turgidum*. *Archives of Microbiology* 202:1563-1569. 10.1007/s00203-020-01859-2

- Erler R, Wichels A, Heinemeyer E-A, Hauk G, Hippelein M, Reyes NT, and Gerdt G. 2015. VibrioBase: A MALDI-TOF MS database for fast identification of *Vibrio* spp. that are potentially pathogenic in humans. *Systematic and Applied Microbiology* 38:16-25. 10.1016/j.syapm.2014.10.009
- Freiwald A, and Sauer S. 2009. Phylogenetic classification and identification of bacteria by mass spectrometry. *Nat Protocols* 4:732-742.
- Gao J-l, Sun P, Wang X-m, Lv F-y, Mao X-j, and Sun J-g. 2017. *Rhizobium wenxiniae* sp. nov., an endophytic bacterium isolated from maize root. *International Journal of Systematic and Evolutionary Microbiology* 67:2798-2803. 10.1099/ijsem.0.002025
- Ghyselinck J, Velivelli SLS, Heylen K, O'Herlihy E, Franco J, Rojas M, De Vos P, and Prestwich BD. 2013. Bioprospecting in potato fields in the Central Andean Highlands: Screening of rhizobacteria for plant growth-promoting properties. *Systematic and Applied Microbiology* 36:116-127. 10.1016/j.syapm.2012.11.007
- Gibb S, and Strimmer K. 2012. MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics* 28:2270-2271. 10.1093/bioinformatics/bts447
- Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41:95-98.
- Handelsman J. 2004. Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews* 68:669-685. 10.1128/mmbr.68.4.669-685.2004
- Hoang DT, Chernomor O, von Haeseler A, Minh BQ, and Vinh LS. 2017. UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* 35:518-522. 10.1093/molbev/msx281
- Hornik K, Buchta C, and Zeileis A. 2009. Open-source machine learning: R meets Weka. *Computational Statistics* 24:225-232. 10.1007/s00180-008-0119-7
- Howe AC, Jansson JK, Malfatti SA, Tringe SG, Tiedje JM, and Brown CT. 2014. Tackling soil diversity with the assembly of large, complex metagenomes. *Proceedings of the National Academy of Sciences* 111:4904-4909. 10.1073/pnas.1402564111
- Hsieh TC, Ma KH, and Chao A. 2016. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods in Ecology and Evolution* 7:1451-1456. 10.1111/2041-210x.12613
- Hushek D, and Witzel K. 2019. Rapid dereplication of microbial isolates using matrix-assisted laser desorption ionization time-of-flight mass spectrometry: A mini-review. *Journal of Advanced Research* 19:99-104. 10.1016/j.jare.2019.03.007
- Johnston-Monje D, Mousa WK, Lazarovits G, and Raizada MN. 2014. Impact of swapping soils on the endophytic bacterial communities of pre-domesticated, ancient and modern maize. *BMC Plant Biology* 14:233. 10.1186/s12870-014-0233-3
- Junier T, and Zdobnov EM. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. *Bioinformatics* 26:1669-1670. 10.1093/bioinformatics/btq243

- 612 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, and Jermini LS. 2017.
613 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*
614 14:587-589. 10.1038/nmeth.4285
- 615 Kämpfer P, Busse HJ, Kleinhagauer T, McInroy JA, and Glaeser SP. 2016a. *Sphingobacterium*
616 *zeae* sp. nov., an endophyte of maize. *International Journal of Systematic and*
617 *Evolutionary Microbiology* 66:2643-2649. 10.1099/ijsem.0.001100
- 618 Kämpfer P, Glaeser SP, McInroy JA, and Busse HJ. 2016b. *Nocardioides zeicaulis* sp. nov., an
619 endophyte actinobacterium of maize. *International Journal of Systematic and*
620 *Evolutionary Microbiology* 66:1869-1874. 10.1099/ijsem.0.000959
- 621 Kolton M, Sela N, Elad Y, and Cytryn E. 2013. Comparative genomic analysis indicates that
622 niche adaptation of terrestrial *Flavobacteria* is strongly linked to plant glycan
623 metabolism. *PLoS ONE* 8:76704. 10.1371/journal.pone.0076704
- 624 Kumar P, Dubey RC, and Maheshwari DK. 2012. *Bacillus* strains isolated from rhizosphere
625 showed plant growth promoting and antagonistic activity against phytopathogens.
626 *Microbiological Research* 167:493-499. 10.1016/j.micres.2012.05.002
- 627 Labbé JL, Weston DJ, Dunkirk N, Pelletier DA, and Tuskan GA. 2014. Newly identified helper
628 bacteria stimulate ectomycorrhizal formation in *Populus*. *Frontiers in Plant Science*
629 5:579. 10.3389/fpls.2014.00579
- 630 LaMontagne MG. 2020. Gnotobiotic rhizobacterial isolation plant systems and methods of use
631 thereof. US Patent 10801079, Oct. 13, 2020.
- 632 LaMontagne MG, Shetty T, Gajjar T, Kayyuru C, Sriram S, Zhang C, and Buddharaju P. 2017.
633 HABase: A web-application for the analysis of protein spectra and identification of
634 microbial species. In: Arabnia HR, Tran Q-N, and Yang M, editors. BIOCOMP'17. p 77-
635 78. ISBN: 1-60132-450-2.
- 636 Long HH, Sonntag DG, Schmidt DD, and Baldwin IT. 2010. The structure of the culturable root
637 bacterial endophyte community of *Nicotiana attenuata* is organized by soil composition
638 and host plant ethylene production and perception. *New Phytologist* 185:554-567.
639 10.1111/j.1469-8137.2009.03079.x
- 640 Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K,
641 Caugant DA, Feavers IM, Achtman M, and Spratt BG. 1998. Multilocus sequence typing:
642 A portable approach to the identification of clones within populations of
643 pathogenic microorganisms. *Proceedings of the National Academy of Sciences* 95:3140-
644 3145.
- 645 Mashiane RA, Ezeokoli OT, Adeleke RA, and Bezuidenhout CC. 2017. Metagenomic analyses
646 of bacterial endophytes associated with the phyllosphere of a Bt maize cultivar and its
647 isogenic parental line from South Africa. *World Journal of Microbiology and*
648 *Biotechnology* 33:80. 10.1007/s11274-017-2249-y
- 649 McInroy JA, and Kloepper JW. 1995. Survey of indigenous bacterial endophytes from cotton
650 and sweet corn. *Plant and Soil* 173:337-342. 10.1007/BF00011472
- 651 Mosimann C, Oberhänsli T, Ziegler D, Nassal D, Kandeler E, Boller T, Mäder P, and Thonar C.
652 2017. Tracing of two *Pseudomonas* strains in the root and rhizoplane of maize, as related

653 to their plant growth-promoting effect in contrasting soils. *Frontiers in Microbiology*
654 7:2150. 10.3389/fmicb.2016.02150

655 Mousa WK, Shearer CR, Limay-Rios V, Zhou T, and Raizada MN. 2015. Bacterial endophytes
656 from wild maize suppress *Fusarium graminearum* in modern maize and inhibit mycotoxin
657 accumulation. *Frontiers in Plant Science* 6:805. 10.3389/fpls.2015.00805

658 Murugaiyan J, Lewin A, Kamal E, Bakula Z, van Ingen J, Ulmann V, Unzaga Barañano MJ,
659 Humiecka J, Safianowska A, Roesler UH, and Jagielski T. 2018. MALDI spectra
660 database for rapid discrimination and subtyping of *Mycobacterium kansasii*. *Frontiers in*
661 *Microbiology* 9:587. 10.3389/fmicb.2018.00587

662 Nabti E, Bensidhoum L, Tabli N, Dahel D, Weiss A, Rothballer M, Schmid M, and Hartmann A.
663 2014. Growth stimulation of barley and biocontrol effect on plant pathogenic fungi by a
664 *Cellulosimicrobium* sp. strain isolated from salt-affected rhizosphere soil in northwestern
665 Algeria. *European Journal of Soil Biology* 61:20-26. 10.1016/j.ejsobi.2013.12.008

666 Naveed M, Mitter B, Yousaf S, Pastar M, Afzal M, and Sessitsch A. 2014. The endophyte
667 *Enterobacter* sp. FD17: a maize growth enhancer selected based on rigorous testing of
668 plant beneficial traits and colonization characteristics. *Biology and Fertility of Soils*
669 50:249-262. 10.1007/s00374-013-0854-y

670 Niu B, Paulson JN, Zheng X, and Kolter R. 2017. Simplified and representative bacterial
671 community of maize roots. *Proceedings of the National Academy of Sciences* 114:
672 E2450-E2459. 10.1073/pnas.1616148114

673 Pandey A, Jain R, Sharma A, Dhakar K, Kaira GS, Rahi P, Dhyani A, Pandey N, Adhikari P, and
674 Shouche YS. 2019. 16S rRNA gene sequencing and MALDI-TOF mass spectrometry
675 based comparative assessment and bioprospection of psychrotolerant bacteria isolated
676 from high altitudes under mountain ecosystem. *SN Applied Sciences* 1:278.
677 10.1007/s42452-019-0273-2

678 Park ST, and Kim J. 2016. Trends in next-generation sequencing and a new era for whole
679 genome sequencing. *International Neurolouology Journal* 20:S76-S83.
680 10.5213/inj.1632742.371

681 Petersen LM, Martin IW, Moschetti WE, Kershaw CM, and Tsongalis GJ. 2019. Third-
682 generation sequencing in the clinical laboratory: exploring the advantages and challenges
683 of nanopore sequencing. *Journal of Clinical Microbiology* 58:01315-01319.
684 10.1128/jcm.01315-19

685 Pruesse E, Peplies J, and Glöckner FO. 2012. SINA: Accurate high-throughput multiple
686 sequence alignment of ribosomal RNA genes. *Bioinformatics* 28:1823-1829.
687 10.1093/bioinformatics/bts252

688 Raaijmakers JM, and Mazzola M. 2012. Diversity and natural functions of antibiotics produced
689 by beneficial and plant pathogenic bacteria. *Annual Review of Phytopathology* 50:403-
690 424. doi:10.1146/annurev-phyto-081211-172908

691 Rau J, Eisenberg T, Männig A, Wind, C., Lasch P, and Sting R. 2016. MALDI-UP – an internet
692 platform for the exchange of MALDI-TOF mass spectra. *eJournal* 1:2-17.

- Rijavec T, Lapanje A, Dermastia M, and Rupnik M. 2007. Isolation of bacterial endophytes from germinated maize kernels. *Can J Microbiol* 53:802-808. 10.1139/w07-048
- Rocha I, Ma Y, Souza-Alonso P, Vosátka M, Freitas H, and Oliveira RS. 2019. Seed coating: a tool for delivering beneficial microbes to agricultural crops. *Frontiers in Plant Science* 10. 10.3389/fpls.2019.01357
- Salipante SJ, SenGupta DJ, Cummings LA, Land TA, Hoogestraat DR, and Cookson BT. 2015. Application of whole-genome sequencing for bacterial strain typing in molecular epidemiology. *Journal of Clinical Microbiology* 53:1072-1079. 10.1128/JCM.03385-14
- Sauer S, Freiwald A, Maier T, Kube M, Reinhardt R, Kostrzewa M, and Geider K. 2008. Classification and identification of bacteria by mass spectrometry and computational analysis. *PLoS ONE* 3:2843. 10.1371/journal.pone.0002843
- Sboner A, Mu XJ, Greenbaum D, Auerbach RK, and Gerstein MB. 2011. The real cost of sequencing: higher than you think! *Genome Biology* 12:125. 10.1186/gb-2011-12-8-125
- Seuylemezian A, Aronson HS, Tan J, Lin M, Schubert W, and Vaishampayan P. 2018. Development of a custom MALDI-TOF MS database for species-level identification of bacterial isolates collected from spacecraft and associated surfaces. *Frontiers in Microbiology* 9:780. 10.3389/fmicb.2018.00780
- Shehata HR, Lyons EM, Jordan KS, and Raizada MN. 2016. Bacterial endophytes from wild and ancient maize are able to suppress the fungal pathogen *Sclerotinia homoeocarpa*. *Journal of Applied Microbiology* 120:756-769. 10.1111/jam.13050
- Singhal N, Kumar M, Kanaujia PK, and Viridi JS. 2015. MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Frontiers in Microbiology* 6:791. 10.3389/fmicb.2015.00791
- Song W, Pijanowski BC, and Tayyebi A. 2015. Urban expansion and its consumption of high-quality farmland in Beijing, China. *Ecological Indicators* 54:60-70. 10.1016/j.ecolind.2015.02.015
- Strejcek M, Smrhova T, Junkova P, and Uhlik O. 2018. Whole-cell MALDI-TOF MS versus 16S rRNA gene analysis for identification and dereplication of recurrent bacterial isolates. *Frontiers in Microbiology* 9:1294. 10.3389/fmicb.2018.01294
- Suzuki R, and Shimodaira H. 2006. Pvcust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22:1540-1542. 10.1093/bioinformatics/btl117
- Tan Z, Hurek T, Gyaneshwar P, Ladha JK, and Reinhold-Hurek B. 2001. Novel endophytes of rice form a taxonomically distinct subgroup of *Serratia marcescens*. *Systematic and Applied Microbiology* 24:245-251. 10.1078/0723-2020-00002
- Tanaka T, Kawasaki K, Daimon S, Kitagawa W, Yamamoto K, Tamaki H, Tanaka M, Nakatsu CH, and Kamagata Y. 2014. A hidden pitfall in agar media preparation undermines cultivability of microorganisms. *Applied and Environmental Microbiology* 80:7659-7666. 10.1128/aem.02741-14
- Timperio AM, Gorrasi S, Zolla L, and Fenice M. 2017. Evaluation of MALDI-TOF mass spectrometry and MALDI BioTyper in comparison to 16S rDNA sequencing for the

733 identification of bacteria isolated from Arctic sea water. *PLoS ONE* 12:0181860.
734 10.1371/journal.pone.0181860

735 Tkacz A, and Poole P. 2015. Role of root microbiota in plant productivity. *Journal of*
736 *Experimental Botany* 66:2167-2175. 10.1093/jxb/erv157

737 Tomachewski D, Galvão CW, de Campos Júnior A, Guimarães AM, Ferreira da Rocha JC, and
738 Etto RM. 2018. Ribopeaks: a web tool for bacterial classification through m/z data from
739 ribosomal proteins. *Bioinformatics* 34:3058-3060. 10.1093/bioinformatics/bty215

740 Trenberth KE, Dai A, van der Schrier G, Jones PD, Barichivich J, Briffa KR, and Sheffield J.
741 2014. Global warming and changes in drought. *Nature Climate Change* 4:17-22.
742 10.1038/nclimate2067

743 Trifinopoulos J, Nguyen L-T, von Haeseler A, and Minh BQ. 2016. W-IQ-TREE: a fast online
744 phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research* 44:W232-
745 W235. 10.1093/nar/gkw256

746 Tsang AKL, Lee HH, Yiu S-M, Lau SKP, and Woo PCY. 2017. Failure of phylogeny inferred
747 from multilocus sequence typing to represent bacterial phylogeny. *Scientific Reports*
748 7:4536. 10.1038/s41598-017-04707-4

749 Uhlik O, Strejcek M, Junkova P, Sanda M, Hroudova M, Vlcek C, Mackova M, and Macek T.
750 2011. Matrix-Assisted Laser Desorption Ionization (MALDI)-Time of Flight Mass
751 Spectrometry- and MALDI Biotyper-based identification of cultured biphenyl-
752 metabolizing bacteria from contaminated horseradish rhizosphere soil. *Applied and*
753 *Environmental Microbiology* 77:6858-6866. 10.1128/aem.05465-11

754 Wang X-Q, Bi T, Li X-D, Zhang L-Q, and Lu S-E. 2015. First report of corn whorl rot caused by
755 *Serratia marcescens* in China. *Journal of Phytopathology* 163:1059-1063.
756 10.1111/jph.12366

757 Wang X, Qu C, Wang W, Zheng Z, Liu F, An M, and Miao J. 2019. Complete genome sequence
758 of marine *Bacillus* sp. Y-01, isolated from the plastics contamination in the Yellow Sea.
759 *Marine Genomics* 43:72-74. 10.1016/j.margen.2018.05.002

760 Weis CV, Jutzeler CR, and Borgwardt K. 2020. Machine learning for microbial identification
761 and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic
762 review. *Clinical Microbiology and Infection* 26:1310-1317. 10.1016/j.cmi.2020.03.014

763 Wright ES, Yilmaz LS, and Noguera DR. 2012. DECIPHER, a search-based approach to
764 chimera identification for 16S rRNA sequences. *Applied and Environmental*
765 *Microbiology* 78:717-725. 10.1128/aem.06516-11

Figure Legends

Figure 1. Cosine versus Jaccard similarities for pairwise comparisons. Line illustrates a mathematical model of the hyperbolic relationship described in Methods, not a fit to the data.

Figure 2. Cosine similarity versus 16S identity. Cosine similarity was weighted with values predicted from Jaccard coefficients using the hyperbolic model in Figure 1 as described in Methods. Line represents a sigmoidal curve fit.

Figure 3. Rarefaction analysis of library generated from the maize microbiome. X-axis presents number of isolates in the library (filled circles) or an extrapolation from the existing library (open circles). Confidence intervals present standard error bars. Figures were generated with the default parameters in iNext. Top) Y-axis presents number of taxonomic units defined by similarity of mass spectra. Bottom) Y-axis presents fraction of total diversity recovered.

Figure 4. Phenogram generated from 16S rRNA gene sequences. Partial sequences were aligned against curated sequences with SINA. ModelFinder was used select the appropriate phylogenetic maximum likelihood model within IQ-TREE. Bootstrap values were calculated with UFBoot2. Genbank accession numbers and species identifications are provided for reference sequences. Scale indicates substitutions, where a distance of 0.04 corresponds to 96% identity.

Figure 5. Hierarchical clustering of mass spectra for isolates reliably identified with Bruker MALDI Biotyper system. AU and BP values present approximately unbiased probability values and bootstrap probabilities assigned with pvclust. MTUs and species identification given by the Biotyper system are indicated at each node. Height presents dissimilarity between nodes.

Figure 6. MTUs detected in library generated from the rhizoplane and endosphere of maize. The number of isolates clustered in each MTU is indicated. Singletons are not shown.

Figure S1. Jaccard similarities for pairwise comparisons of peak shared between mass spectra versus number of iterations of alignments. Y-axis presents number of peaks shared. Points correspond to values calculated with different iterations of randomly selected parameters used to align spectra (see Methods).

Figure S2. Jaccard similarities for pairwise comparisons of peak shared between mass spectra versus peaks detect. Y-axis presents number of peaks shared. X-axis presents peaks. Points correspond to values calculated with different iterations of randomly selected parameters used to align spectra (see Methods).

Figure S3. MTUs detected in library generated from the two varieties of maize. The number of isolates clustered in each MTU is indicated. Singletons are not shown.

Distance: euclidean
Cluster method: ward.D2