Vigilada Mineducación

Compositionally-aware bioinformatic workflow for the analysis of *16S* gene sequencing variants of Banana (*Musa acuminata*) plant endophytes.

Luis Alfonso Arteaga Figueroa
Trabajo de grado

Javier Correa Álvarez, PhD.

UNIVERSIDAD EAFIT

Escuela de Ciencias
Biología
Medellin
2021

Compositionally-aware bioinformatic workflow for the analysis of *16S* gene sequencing variants of Banana (*Musa acuminata*) plant endophytes.

Asesor: Prof. Javier Correa Álvarez **Luis Alfonso Arteaga-Figueroa**

Abstract: Marker gene sequencing of the *16S* has become the gold standard to characterize microbial communities from all environments. One of those environments are plants, which are in constant interaction with microbes that have the ability to colonize their surface and interior and perform very valuable metabolic functions. In the present study, we assessed different tools and databases to study *16S* marker sequencing data considering distinct biases of the technique in order to unravel the endophytic dynamics of Banana (*Musa acuminata*) under different conditions such as the infection by *Pseudocercospora fijiensis* and the inoculation of *Bacillus subtilis* EA-CB0575.

1 Introduction

Microbial communities are present through most of the known ecosystems of the Earth, these communities, also called microbiota along with their genomes the microbiome, play crucial roles in sustaining life as we know it today [1, 2]. Thus, its characterization is vital to understand their roles in the ecosystem [3]. Many techniques for characterizing microbial communities have emerged over time, such as: cell culture, DDGE, FISH or T-RLFPs [3, 4], but none of these techniques has grown as popular as marker gene sequencing [4].

Marker gene sequencing to characterize microbial communities has been done traditionally using the *16S* ribosomal RNA (*16S*) ribosomal marker. The 16S gene was first used for phylogenetic purposes back in 1985, due to its nature of containing regions of both conservation and variability, which ease the primer design over the conserved regions and the identification of organisms based in the hypervariable regions [5]. Subsequently, *16S* became the most used marker gene for profiling microbial communities; but it wasn't until the development of High-throughput sequencing (HTS) technologies that this technique became affordable and routine [3]. Thus, amplifying and sequencing *16S* has become one of the standard techniques to survey and compare microbial communities across space, time and environments [6]. Nevertheless, implementing HTS carried with it some trade-offs [7].

The issues of HTS arise due to the fragmentary and compositional nature of sequencing data, the former issue is related to the read length supported by the platforms, while the latter is related to the nature of sampling during the sequencing. Most "Second generation" or "Next generation" sequencing (NGS) methods output reads between 200 and 600 basepairs (bps) while full-length *16S* gene is around 1500 bps long [6]; hence, only a portion of *16S* can be sequenced by NGS methods and this represents a problem for taxonomic classification and phylogenetic reconstruction, which directly impacts in the calculation of several α-diversity and β-diversity metrics [8]. Additionally, primer election to amplify the targeted region to sequence introduces the "PCR bias", in which variations over the conserved region where primers are designed result in unequal amplification of the *16S* over the sample [9], this trade-off is highly relevant for plant associated microbiota studies as universal *16S* primers also result in plastid and mitochondrial derived amplicons [10].

Among the mentioned issues of HTS, we find the compositional bias, which is due to the limited capacity of the sequencing instruments, that can only deliver a given number of reads, inevitably, the total read count of an HTS run is of a fixed-size, which represents a random sample of the molecules in the sampled ecosystem. This directly affects the captured abundance of sampled organisms, which subsequently affects β-diversity calculations [7, 11]. Additionally, related to the compositional bias, we find Esparcity, which is the large amount of zeros in the count matrices of data derived

from HTS experiments [11].

Furthermore, with the rise of marker gene HTS-based characterization of microbial communities, also came the realization that a vast portion of Earth' microbiota is uncultured, therefore uncharacterized [12]. This had led to a quest to quantify how much of certain communities remains uncultured, which members need an extra effort to be cultured and how representative the isolates are of the sequenced data [13]. Nevertheless, some considerations as gapped distances metrics [14] must be considered to accurately characterize this gap.

Although the framework of amplicon analysis was established in the past decade, analysis methods and standards have been evolving rapidly over the past few years, the quick accumulation of new methods makes challenging for researchers to choose suitable software and pipelines for data analysis and interpretation [4]. One of the recently published pipelines is for microbiome analysis is Qiime2 (q2), which includes plugins of the latest-generation tools for quality check, taxonomy assignment and phylogenetic reconstruction [15].

Plants are in constant communication with a wide variety of microorganisms in its environment [16], microbial habitats in plants comprise the bulk soil, phyllosphere, spermosphere, rhizosphere, endosphere [17]. Endophytes, which inhabit the endosphere, are defined as those microorganisms that colonize internal plant tissue without causing harm to the plant. These endophytes can promote plant growth, elicit defense or protect against pathogens and act as remediators of abiotic stress [16], these traits make endophytes potential natural resources to control abiotic and biotic stresses [18].

In the present work, we assessed different tools and databases for the analysis of *16S* marker gene sequencing data derived from both HTS sequencing and amplicons from cell cultures, in (mostly) Qiime2. Thus, the present work designed a state-of-the-art analysis workflow, necessary to unravel the dynamics of the endophytic microbiota of two varieties of Banana (*Musa acuminata*) var. Calcutta-4 and Williams under different treatments PGPR *Bacillus subtilis* EA-CB0575 [19] inoculation, *Pseudocercospora fijiensis* infection and plants without *P. fijiensis* or *B. subtilis* EA-CB0575 in order to identify promising endophytes for future bioproducts and study the effect of treatments on the banana microbiota.

2 Methods

2.1 Previous work.

Wetlab experiments such as *B. subtilis* inoculation, *P. fijiensis* infection, greenhouse plant growth, DNA extraction and *16S* amplification was performed by Luisa F. Posada-Uribe in unpublished work. Then facility sequencing by BaseClear (Nederlands), of the hypervariable V3-V4 region (primers 341F/805R), in combination with chloroplast and mitochondrial blocking primers, was performed. Sanger reads obtained in previous work by Adarve-Rengifo (2020) [20], were basecalled using Tracy v0.5.9 [21], then trimmed with Cutadapt [22] v2.8 and finally converted to a Qiime2 artifact for downstream analyses such as taxonomy assignment.

2.2 Wetlab experiments

Fifteen g of Calcutta 4 and Williams banana, infected with *P. fijiensis*, inoculated with B. subtilis EA-CB0575 plants sections (leaves, roots or pseudostem) and neither inoculated and infected, were disinfected and macerated using liquid N2 a for DNA extraction (DNeasy PowerSoil Kit, Qiagen) according to protocol recommended by manufacturer with complementary steps for DNA concentration. After discarding spin filter 4 mL NaCl 5 M were added and inverted, 200 mL of cold ethanol 96% were added and tubes were inverted and centrifuged at 11500 rpm during 5 min at root temperature and supernatant was discarded. Finally, residual ethanol was discarded using Vacuum Concentrator (mode V-HV, 30°C during 20 min) and pellet was resuspended on sterile Tris-HCl 10mM pH 8.5 and storage at -80°C. DNA concentration was evaluated by gel electrophoresis 1% agar, Nanodrop™ spectrophotometer (Thermo Fisher Scientific, USA) and Qubit® 2.0 Fluorometer (IThermoScientific).

DNA samples with concentration greater than 10 ng/mL were submitted to

2.3 Bioinformatics analysis.

2.3.1 Quality check and Amplicon Sequence Variant inference

Raw data was quality checked with FastQC and MultiQC [23], then *16S* rRNA amplicon sequence variants (ASVs) were inferred using DADA2 pipeline in Qiime2 v2020.08 [24]; forward and reverse reads were truncated at 280 and 250 bp respectively, the 10 first bases were trimmed for all reads, pseudo-pooling method was set on and all other parameters were default.

2.3.2 Database curation and Taxonomy assignment

Database curation was done with q2 plugin RESCRIPt [25] over SILVA v138 [26, 27], Bacteria GTDB r220 [28], GreenGenes 13_8 [29] and NCBI Refseq [30]. Taxonomic assignment was performed with different approaches and databases in order to compare, such as naïve bayes (NB) classifiers [31], VSEARCH [32] and BLAST[33], all implemented in q2 feature-classifier [34]. Briefly, for taxonomic assignment with naïve bayes classifiers, first primer specific databases were extracted with q2 extract-reads [35] for primer sets 341F-805R and 8F-907R, then NB classifiers were trained using q2 plugin fit-classifier-naive-bayes, and finally ASVs were classified using q2 plugin classify-sklearn. For VSEARCH classification, ASVs were compared against primer specific databases. Comparisons among different databases were made with q2-rescript evaluate-taxonomy. ASVs classified either as Mitochondria or Chloroplast were removed with q2-filter-table. As long for culture amplicon classification was performed with 8F-907R primer specific databases.

2.3.3 Phylogenetic analysis

Phylogenetic analyses were performed de novo and by Saté-enabled phylogenetic placement (SEPP). First, ASVs were aligned with q2-mafft [36] then masked for gappy regions, de novo phylogenetic analysis was performed by RAxML [37] with 10 independent replicates. SEPP analysis was performed with q2 fragment insertion [8] default parameters using SILVA 128.1 reference.

2.3.4 α-diversity and β-diversity metrics calculation

Diversity indexes were calculated with q2-core metrics, q2-alpha and q2-beta plugins. As for phylogenetic-aware metrics such as Faith phylogenetic diversity, two analyses were computed independently modifying its input tree based in the methodology of reconstruction, and a rarefaction depth of 9200 counts where needed [38]. α-diversity was estimated using Shannon as Richness and Evenness estimator, Simpson E and Lladser [39] as Evenness estimators and Fisher-α and Faith's phylogenetic as richness estimators as suggested by Hagerty *et al.* (2020) [40]. Metrics were calculated in q2-alpha and tested using Kruskal-Wallis test [41] implemented in q2-alpha-significance plugin. β-diversity was estimated with Jaccard, Bray-Curtis, and DEICODE Aitchinson distance indexes with q2 diversity beta and q2 DEICODE [42], then a Principal coordinate analysis (PCoA) was performed for all beta diversity indexes assessed, then DEICODE Aitchinson distance was tested using Permutational multivariate analysis of variance (PERMANOVA) [43], and Analysis of similarities (ANOSIM), QZA distance matrix artifacts were imported into R with qiime2R for visualization.

2.3.5 Differential abundance

Differentially abundant ASVs for the three variables independently, among contrast of treatments, were calculated using an agglomerative approach implemented in rANOMALY [44] for non-compositional methods (DEseq2, Metagenomeseq and Metacoder) and q2-composition (ANCOM v1) and ANCOM-BC [45] for compositional methods, all with default parameters. Additionally, as ANCOM

is affected by the inacuracies of automatic taxonomy assignment when counts are pooled at given taxonomic level, we performed a comparison between ANCOM results at genus level between SILVA v138.1 and GTDB r220, genera were considered differentially abundant when q-value < 0.05.

2.3.6 Cultured collection representation of NGS sequencing data

First, regions belonging to the V3-V4 region of the basecalled chromatograms from the culture collection were extracted using q2 extract-reads and 341F-805R primer set. Resulting sequences were then used as the database input for VSEARCH search and ASVs representative sequences as query, we performed searches with the Marine Biological Lab identity definition, which takes gap openings into account, as suggested by [14]. An ASV was considered represented by the culture collection when VSEARCH reported a match with ≥ 0.97 of identity score with a strain sequence. Then, features from each category (represented or not) were collapsed with q2 feature-table group by the sum of its frequency by treatment. Thus, proportion of the *16S* sequencing data represented within the culture collection was calculated by dividing the sum of frequency of represented ASVs and the total frequency for each treatment.

3 Results

48 libraries were sent for facility sequencing but only 36 were successfully sequenced, Calcutta-4 was the treatment most affected by technical failures (10 of the samples belonged to this treatment). Nevertheless, other factors such as plant organ and PGPR/infection remained evenly distributed. After ASV inference and plastid/mitochondrial sequence removal, 9941 ASVs were kept for downstream analyses.

3.1 Taxonomy assignment

Metrics for the assessment of the tested databases are available at Figure 1, SILVA and GTDB performed best over NCBI RefSeq and GreenGenes. Results for general taxonomy assignment at phylum level are available in Figure 2. At phylum level, Proteobacteria was the most abundant, followed by Bacteroidota, Actinobacteriota, Firmicutes and Acidobacteriota. Nevertheless, an unprecedent abundance of recently described taxa such as Patescibacteria (CPR), PVC (Planctomycetota, Verrucomicrobia, Chlamydiota) and Gemmatimonodota was detected (Figure 2). The higher diversity of classified taxa is found at genus level, which has the highest count of Unique labels and Taxonomic Entropy, therefore, differentially abundant taxa was calculated with this level for both taxonomy assignments (Figure 1). Taxonomic classification and strain count from the cell culture collection is available at Figure 3.

3.2 α-diversity and β-diversity metrics

3.2.1 α-diversity

Boxplots for α-diversity visualization are available at Figure 4. Overall, most metrics are significantly different when compared in the Organ (Aerial vs Roots) Factor, samples from roots have significantly higher richness according to the evaluated metrics. Evenness metrics such as Lladser and Simpson E are also significantly different in this treatment. Furthermore, evaluated richness metrics are significantly different when *P. fijiensis* infection and *B. subtilis* EA-CB0575 are compared against the control. Nevertheless, none of the evaluated metrics are significantly different for the variety treatment.

3.2.2 β-diversity

PCoA of the different β-diversity metrics assessed are available at Figure 5. All β-diversity metrics assessed showed a clear and significant differentiation of Roots and Aerial samples (Aitchinson-DEICODE, PERMANOVA and ANOSIM p-value: 0.001). Nevertheless, none of the β-diversity metrics assessed showed a clear differentiation be-

tween treatments of the PGPR/pathogen factor (Aitchinson-DEICODE, PERMANOVA p-value: 0.05 and ANOSIM p-value: 0.11). Additionally, Aitchinson-DEICODE and Bray-Curtis distances showed a clear differentiation between Calcutta-4 and Williams, specific to Root samples, however it is the only metric in which this differentiation is observed and is significant (Aitchinson-DEICODE, PERMANOVA and ANOSIM p-value: 0.001). The top variants contributing variation belong to *Acidocella*, *Streptomyces*, *Cutibacterium*, *Sphingomonas*, *Trinickia*, *Dyella* and members of the Rhizobiaceae and Micrococcaceae families.

3.3 Differential abundance

Relative abundance of taxa determined as differentially abundant (DA) by ANCOM-BC between all factors is available at Figure 6. Although characterizing the differences between the rhizosphere and phyllosphere are outside the scope of this study, ANCOM-BC identified at least 60 genera DA with both taxonomies in the Organ Factor, equally distributed between factors. Q2-composition ANCOM and Q2-Aldex2 failed to output taxa DA in the root treatment, nevertheless all the taxa identified as DA in the Aerial treatment for both plugins were also identified as DA by ANCOM-BC. To synthesize this variation the top five DA (Higher W value) genera of ANCOM-BC–GTDB results were selected. For the root treatment the top five genera are *Streptomyces*, *Terracidiphilus*, *Devosia*_A, CAIPFQ01 and 2011-GWC2-44-17. As for leaves, the five top are *Methylobacterium*, *Staphylococcus*, *Actinomycetospora*, *Corynebacterium* and *Lawsonella*. Finally, a distinct enrichment of families Chitinophagaceae, GWC2-71-9 and UBA2163 families was observed in the roots, in which 6 taxa are DA for Chitinophagaceae (*Chitinophaga, Niastella, Flavipsychrobacter, Paraflavitalea, Sediminibacterium* and NS-102), two taxa for GWC2-71-9 (DSTP01 and JABFSM01) and two for UBA2163 (UBA10103 and C7867-001).

Meanwhile for the variety treatment and the PGPR/pathogen factor a high level of variation between analyses was found, therefore, this makes more difficult to analyze which taxa is DA. Initially *Kosakonia* was consistenly determined as DA between varieties (specifically in Calcutta-4 variety) by compositional and non-compositional DA methods, as well as, *Noviherbaspirillum* (DA in Calcutta-4) and *Acidocella* (DA in Williams), in a lesser extent. Beyond this consistency, there is not overlap between taxa determined as DA by ANCOM-BC when results for both taxonomy assignments are compared. Other taxa determined as DA by ANCOM-BC with GTDB classification are *Sphingomonas*, *Methyloceanibacter*, AC-14 and VKM-B-2647 which are more abundant in Calcutta-4 variety along *Acidobacterium* and *Shinella* which are more abundant in Williams variety. Additionally, taxa determined as DA by ANCOM-BC with SILVA taxonomy are *Niabella*, *Haoranjiania* (both from Chitinophagaceae), *Thiomonas*, Possible genus 04 (Fibrobacteriaceae), *Ca. Jorgensenbacteria* and *Ca. Brownia*.

Furthermore, *Crossiella* is the only Genus consistently DA in samples without PGPR inoculation or infection across most of the analyses. Beyond this consistency, again, there is not overlap between taxa determined as DA when results are compared by taxonomy assignments. Taxa determined as DA by ANCOM-BC and GTDB taxonomy are *Methylotenera*, C7867-002, SHVA01 and Pan189 DA in *P. fijiensis* infection, *Gemmatimonas* DA in samples with PGPR application and F1-20-MAGs016, AC-14 DA in samples without PGPR inoculation or infection. On the other hand, taxa determined as DA by ANCOMBC and SILVA taxonomy are *Dechloromonas*, *Terrimonas*, *Ca. Brownia*, env.OPS.17 and uncultured members of Pedosphaeraceae DA in *P. fijiensis* infection, Subgroup 13 DA in samples with PGPR and *Arthrobacter* and uncultured members of Methyloligellaceae DA in samples without PGPR inoculation or infection.

To summarize, taxa determined as differentially abundant by ANCOM-BC in combination with GTDB r220 are spread through phyla: Proteobacteria which taxa is mainly DA in the variety treatment, Actinobacteri-

ota which is characterized by an increase of AC-14 taxon in Calcutta 4 and plants without PGPR or infection and a distinct increase of this group in leaves with 7 DA genera for this treatment; Acidobacteriota with DA taxa in Williams (Acidobacterium) and roots, Firmicutes with DA taxa in leaves and Calcutta-4 (VKM-B2647), Patescibacteria with DA taxa in infected plants (C7867-002) and roots, Gemmatimonadota with a Genus DA in plants with PGPR application (Gemmatimonas) and PVC superphylum, in which Chlamydiota taxon is DA in roots and Planctomycetota taxa DA either in infected plants or plants without infection or PGPR application.

Finally, it is worthy to mention that taxa determined as DA by ANCOM-BC in combination with SILVA was highly concordant with results from non-compositional methods assessed with rANOMALY, in opposition to results of ANCOM-BC in combination with GTDB which are hardly correlated with non-compositional methods except for previously mentioned cases.

3.4 Culture Collection representation of NGS data

Cultivated organisms belong to 57 genera, which represent roughly the 5% of genus categories detected in NGS data. Differentially abundant genera included in the cultured collection include: *Methylobacterium*, *Curtobacterium*, *Brevibacterium*, *Streptomyces*, *Sphingomonas* and *Kosakonia*. Nevertheless, the proportion of NGS data represented in the culture collection by treatment presented interesting variation: As for variety treatment, the proportion is 0.37 for Calcutta-4 and 0.35 for Williams, as long for Organ, the proportion in Roots is 0.22 and Aerial is 0.53, finally for the proportion for the PGPR inoculation/infection treatment is 0.27 for the infection, 0.3 for the PGPR inoculation and 0.52 for the microbiome of plants without infection or PGPR inoculation. As for features matching *B. subtilis* EA-CB0575 at 0.996 identity score the proportion was of 0.002 in plants without infection or inoculation, 0.0007 for plants with PGPR inoculation and 0.002 in infected plants.

4 Discussion

4.1 Diversity metrics

Richness metrics were significantly different between plants without infection/PGPR application and infected plants, this result is quite contrasting with previous studies as infection by *Fusarium oxysporum sub. cubensis* 4 (Foc4) does not correlate with variations observed in the assessed α-diversity metrics [46]. Furthermore, Aitchinson-DEICODE and Bray-Curtis distances were able to differentiate between Calcutta-4 and Williams in some extent, revealing that microbiomes of these two varieties differ mostly in the roots. Nevertheless, to clarify both tendencies in the assessed diversity metrics further research is needed.

4.2 Differentially abundant taxa

As one of the principal aims of the project is to identify promising endophytes for future bioproducts, the identification of taxa varying across our co-variates is crucial. Differentially abundant groups across all factors display an unprecedent yet specific variation. To illustrate this, Methanol is one of the most abundant compounds in plants as a result of its byproduct during plant growth and decay [47] and it has been recognized the phyllosphere, rhizosphere and endosphere as suitable habitat for methylotrophic communities (Organisms capable of metabolizing C1 compounds)[48]. In this study we found a striking variation of known and potential methylotrophs across our factors, including: *Methylotenera* [49], *Methyloceanibacter* [50], *Methylobacterium* [51] and the phylum Planctomycetota [52].

Methylotrophs in association with the plant can promote plant growth as they perform crucial biological functions such as Nitrogen-fixation, P, K and Zn solubilization, production of hormones and Fe-chelating compounds or indirect biological control [53]. Just *Methylobacterium* alone has potential to act as bio-estimulant, bio-fertilizer and bio-controller

[51]. Nevertheless, the role of other methylotrophs remains unknown, such as *Methylotenera* or Pan189 during the infection by *P. fijiensis* or *Methyloceanibacter* colonizing differentially Calcutta-4. Although, metabolic potential of Planctomycetes taxa DA remains to be assessed by other methods such as Cell culture or Metagenomics, it is clear the lack of knowledge regarding the role of this phylum as endophyte, as part of the methyltrophic network, and during the infection of *P. fijiensis.* Furthermore, Pan189 taxon is currently affiliated to the Planctomycetaceae family which members are known to perform annamox (the anaerobic oxidation of ammonium) [52]. Even without potential metabolic features previously mentioned, Planctomycetes represent a potential source for small bioactive molecules and enantioselective enzymes [54].

Known Plant Growth Promoting taxa (*Kosakonia*, *Noviherbaspirillum*, *Sphingomonas*, *Shinella*) are also DA in the variety factor. Consistently DA in Calcutta-4, *Kosakonia* is both an enigma and on opportunity, as various members of this genus have been found accomplishing many roles, from alleviating the abiotic stress such Salinity, Cold or caused by bioavailable metals [55] to act as control to biotic stress caused by pathogens such as *Fusarium spp.* [56, 18] or *Ralstonia syzygi* [57]. Nevertheless, members of this genus (*Kosakonia cowanii*, specifically) has been also found pathogenic for humans, plants, insects, birds [55]. Further research is needed in order to isolate, culture, characterize and understand the role this genus is fulfilling as endophyte in Calcutta-4 and their biotechnological potential. As long for *Noviherbaspirilum*, some members of this genus have shown potential for denitrification [58], beyond this, its biotechnological potential is unclear. On the other hand, *Sphingomonas* are mainly known for their ability to degrade environmental pollutants but have also been reported to be antagonistic against bacteria and fungi [59] and to produce gibberelins [60].

Furthermore, genera DA en Williams (*Acidocella*, *Acidobacterium* and *Shinella*) beyond the reported plant promoting growth abilities of *Shinella* [61], show interesting biotechnological potential. Both *Acidocella* and *Acidobacterium* have the potential to carry out Fe(III) reduction [62], in which ferric ions are converted to ferrous ions, this is beneficial to the plant as Fe(II) is more soluble [63]. Especially *Acidocella* members which can catalyze this reduction with organic low molecular weight compounds as electron acceptors [64]. Additionally, both *Acidocella* and *Acidobacterium* have recently shown potential for Nitrogen uptake, P and Zn solubilization, Siderophore production and Pyrroloquinoline quinone synthesis for plants growing in Oil-contaminated soils [65].

Meanwhile, *Gemmatimonas* was the only genus differentially abundant in the application of *B. subtilis* EA-CB0575, this result is consistent with previously reported studies which correlated Gemmatimonas with total nitrogen and Foc4 disease suppression after the application of Bio-organic fertilizer [46]. This correlation agrees with previous studies that report that Gemmatimonadota participates in the soil N cycle [66]. Finally, this genus has been also found to be highly abundant in the rhizosphere of healthy sugarcane plants [67]. Nevertheless, as this is a very recently established genus, further research is needed to evaluate the dynamics of this group as endophyte.

Additional to previously mentioned variation of Planctomycetota, we find CAIPFQ01 DA in the roots, taxon that belongs to SM23-39 family (Chlamydia), members this family has the genetic potential to perform acetogenic fermentation and use the arginine deiminase (ADI) pathway to produce ATP [68]. The ecological roles of plant-associated chlamydia are still unknown and matter of ongoing research [69]. Hence, PVC genomic repertoire and physiology remains to be explored in order to clarify its roles as endophytes.

Results from the ANCOM-BC in combination with SILVA are cryptic to analyze as recent changes have been introduced in some of the groups (such as *Arthrobacter* [70]) determined as differentially abundant by this analysis. Besides, many of the variants classified with SILVA changed its taxonomy

when compared to other taxonomy assignments, a very immediate result regards the genus *Dechloromonas* which alleged members are classified as *Azovibrio* or other Rhodocyclaceae genera by GTDB r99 and r202. Thus, it is unclear how accurate these results are.

4.3 Culturable collection in NGS data

Proportion and number of genera of the culturable collection in the NGS data agrees with the previously published studies, in Thomas & Sekhar [71] only 2.6% of detected genera were isolated from endophytes of banana shoots, nevertheless, other studies recover from 20% to 50% of the endophytic microbiota [13], in this study we found that 5% of genera were isolated and NGS is represented in proportion between 0.22 to 0.53, depending on the treatment. Proportion of represented ASVs did not vary in the variety treatment. The highest proportion belonged to the NGS data coming from plants without PGPR inoculation or infection followed by the Aerial organs, this was largely driven by the recovery of *Arthrobacter* and related lineages which are highly abundant in these treatments. Also it is quite clear that isolation efficacies of endophytes from the Roots and infected plants is quite low, further research is needed to cultivate and characterize lineages that might be relevant for the Banana cultivar, from PVC or Gemmatimonadota to *Acidocella* or *Kosakonia.* Furthermore, our understanding of variation in the specific or strain level is limited, as many strains share 100% identity regions in its *16S* sequence and NGS only provides a small fragment to match. This problem has been identified as the principal issue to compare Non-culturable and culturable data sets [13]. This lack of resolution impacts our analysis of the colonization of *B. subtilis* EA-CB0575 as not enough resolution is provided. Further research is needed to assess the different population dynamics of *Bacillus* and the extend of colonization of *B. subtilis* EA-CB0575 inside the plant.

5 Final remarks

In the present study was possible to design and implement a robust analysis methodology to unravel the dynamics of banana endophytes. It was possible to assess different methodologies and databases for taxonomic classification which did impact largely throughout our analysis, specially over the differential abundant taxa determination. It was also possible to determine and correct biases with appropriate diversity metrics and differential abundance tools which took them into account such as Aitchinson-DEICODE and ANCOM-BC, which lead to a higher understanding of the previously mentioned dynamics. Nevertheless, this study is a consequence of its era and falls prey of the pitfalls that it dugs. Extensive amounts of research will be necessary to confirm or deny the results of this study, nevertheless, the considerations laid here for the analysis of *16S* gene marker sequencing data and about the endophytic microbiota of Banana will impact our view of different biological phenomena such as *P. fijiensis* infection and lead to the development of novel bioproducts based in the discussed endophytes that will impact positively the crop.

6 Figures

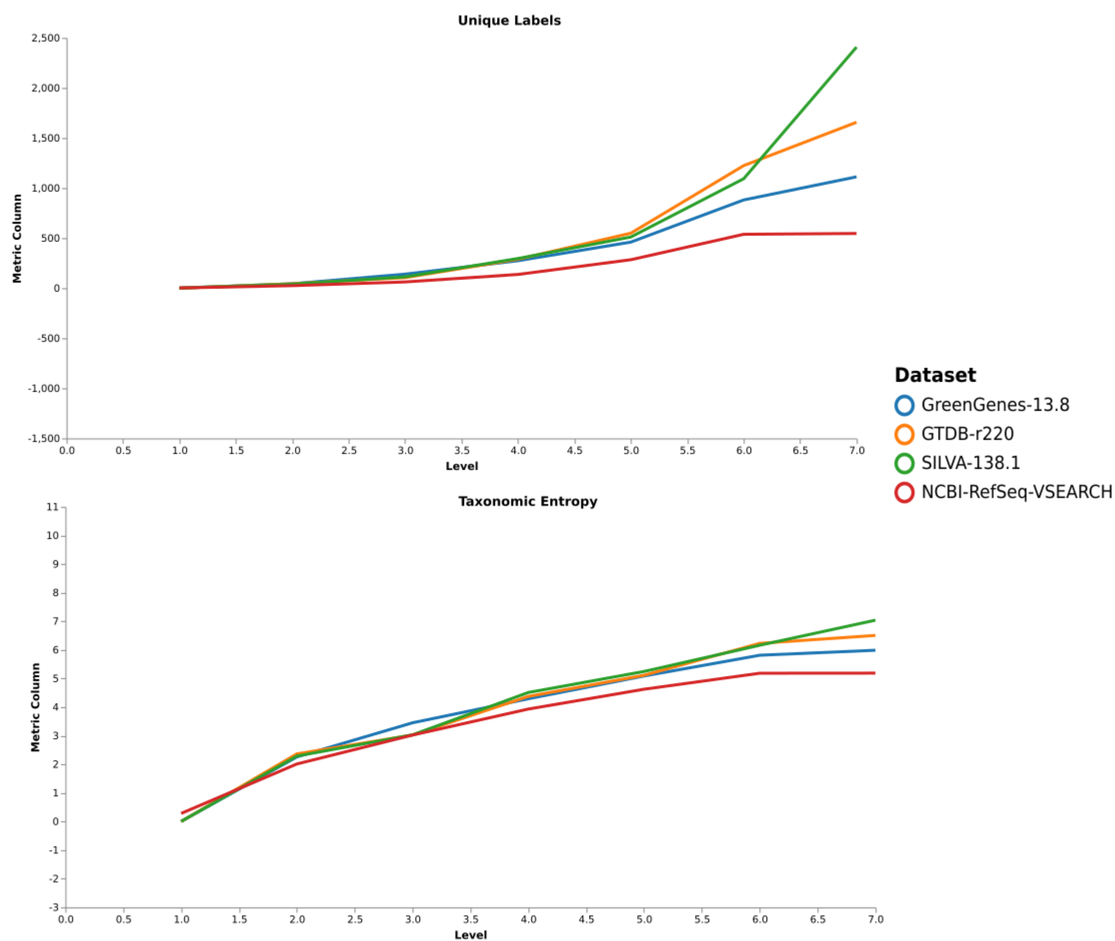

Figure 1: Taxonomic comparisons of assessed databases by Unique labels and Taxonomic entropy, please note that the X-axis is the taxonomic level, 2.0 is the taxonomical phylum level while 6.0 is the genus level

Figure 2: Log10 of frequency of taxa at phylum level with GTDB r220 classification with q2 feature-classifier classify-sklearn

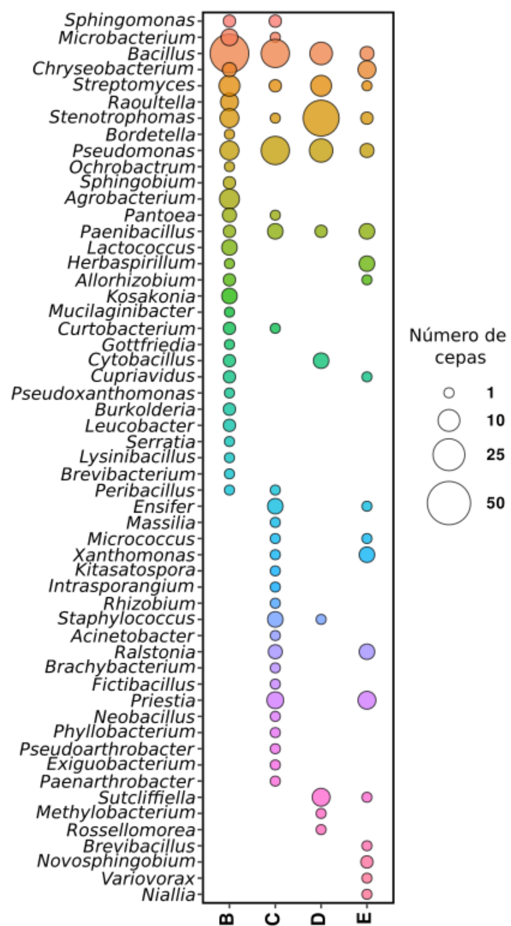

Figure 3: Bubbleplot of the strains collapsed by genera from the cell culture collection, B stands for endophytes of Calcutta-4 *in vitro*, C for isolated endophytes of Williams *in vitro*, D for endophytes of Williams in greenhouse and E for endophytes of Calcutta-4 in greenhouse

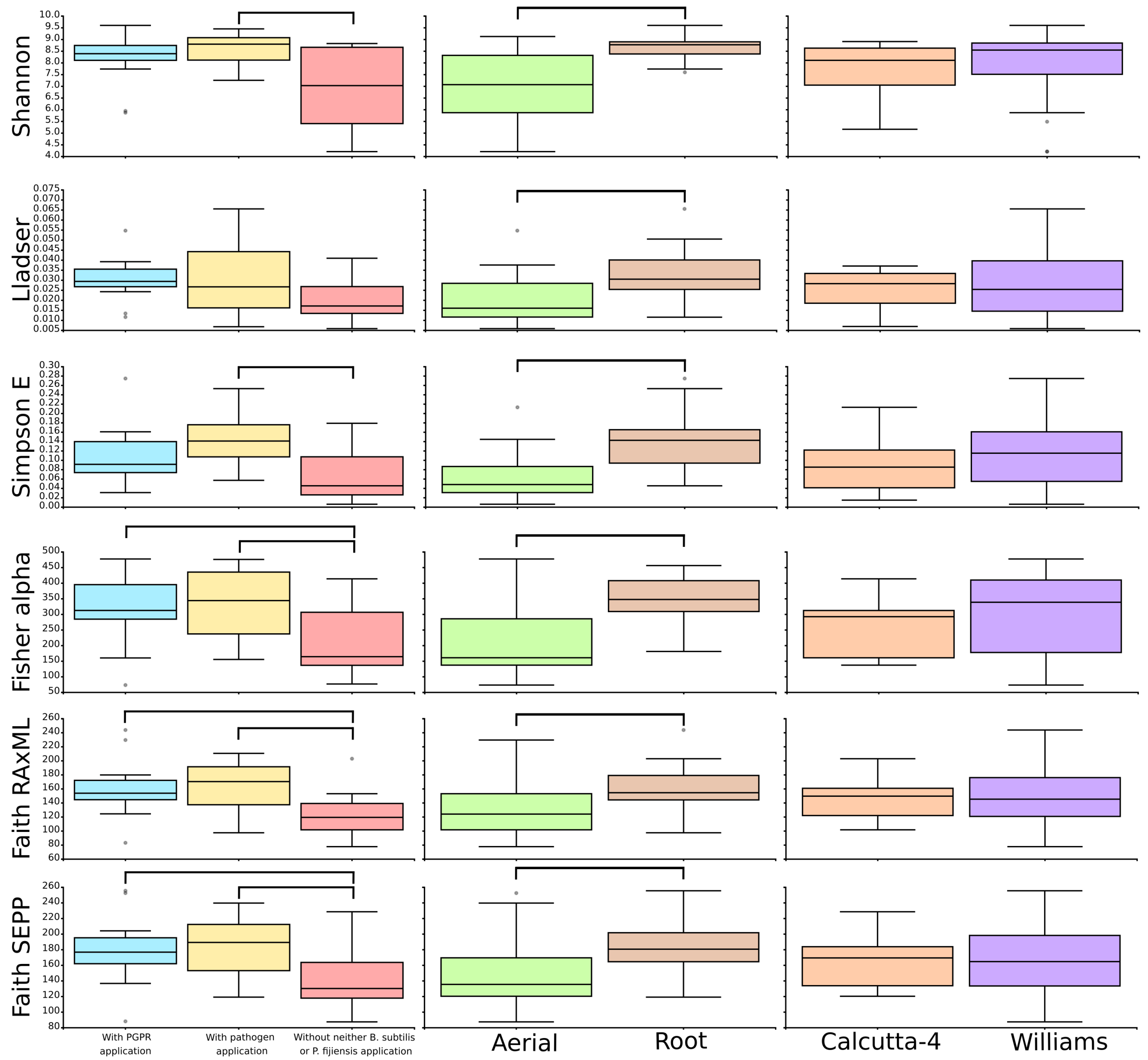

Figure 4: Boxplots of α diversity metrics, bars denoted significant difference with Kruskall-Wallis test among tested treatments

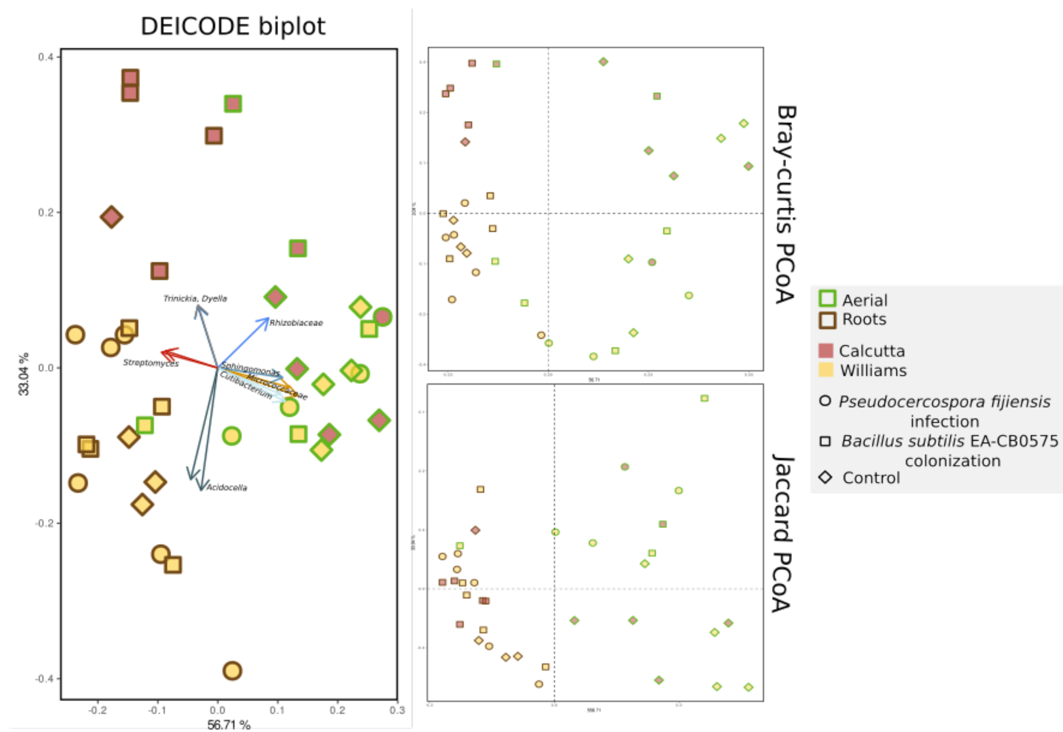

Figure 5: PCoA of assessed β-diversity metrics

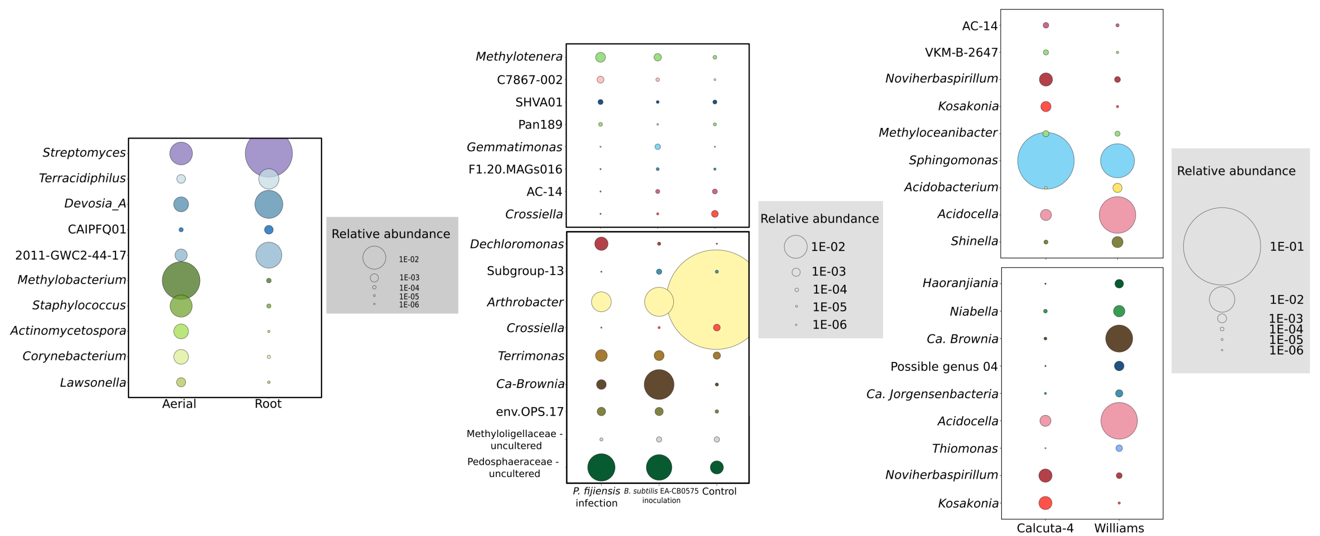

Figure 6: Relative abundance of taxa determined as differentially abundant by ANCOM-BC collapsed by treatment

References

[1] Megan S. Beaudry et al. “Improved Microbial Community Characterization of 16S rRNA via Metagenome Hybridization Capture Enrichment”. In: 12 (Apr. 2021). DOI: 10.3389/fmicb.2021.644662. URL: https://doi.org/10.3389/fmicb.2021.644662.

[2] Silvia Pajares, Brendan J. M. Bohannan, and Valeria Souza. “Editorial: The Role of Microbial Communities in Tropical Ecosystems”. In: 7 (Nov. 2016). DOI: 10.3389/fmicb.2016.01805. URL: https://doi.org/10.3389/fmicb.2016.01805.

[3] Bo Yang, Yong Wang, and Pei-Yuan Qian. “Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis”. In: 17.1 (Mar. 2016). DOI: 10.1186/s12859-016-0992-y. URL: https://doi.org/10.1186/s12859-016-0992-y.

[4] Yong-Xin Liu et al. “A practical guide to amplicon and metagenomic analysis of microbiome data”. In: 12.5 (May 2021), pp. 315–330. DOI: 10.1007/s13238-020-00724-8. URL: https://doi.org/10.1007/s13238-020-00724-8.

[5] Susannah G Tringe and Philip Hugenholtz. “A renaissance for the pioneering 16S rRNA gene”. In: 11.5 (Oct. 2008), pp. 442–446. DOI: 10.1016/j.mib.2008.09.011. URL: https://doi.org/10.1016/j.mib.2008.09.011.

[6] Catherine M. Burke and Aaron E. Darling. “A method for high precision sequencing of near full-length 16S rRNA genes on an Illumina MiSeq”. In: 4 (Sept. 2016), e2492. DOI: 10.7717/peerj.2492. URL: https://doi.org/10.7717/peerj.2492.

[7] Gregory B. Gloor et al. “Microbiome Datasets Are Compositional: And This Is Not Optional”. In: 8 (Nov. 2017). DOI: 10.3389/fmicb.2017.02224. URL: https://doi.org/10.3389/fmicb.2017.02224.

[8] Stefan Janssen et al. “Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information”. In: 3.3 (June 2018). Ed. by Nicholas Chia. DOI: 10.1128/msystems.00021-18. URL: https://doi.org/10.1128/msystems.00021-18.

[9] Isabel Abellan-Schneyder et al. “Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing”. In: 6.1 (Feb. 2021). Ed. by Susannah Green Tringe. DOI: 10.1128/msphere.01202-20. URL: https://doi.org/10.1128/msphere.01202-20.

[10] Rares Lucaciu et al. “A Bioinformatics Guide to Plant Microbiome Analysis”. In: 10 (Oct. 2019). DOI: 10.3389/fpls.2019.01313. URL: https://doi.org/10.3389/fpls.2019.01313.

[11] M. Senthil Kumar et al. “Analysis and correction of compositional bias in sparse sequencing count data”. In: 19.1 (Nov. 2018). DOI: 10.1186/s12864-018-5160-5. URL: https://doi.org/10.1186/s12864-018-5160-5.

[12] Jian-Yu Jiao et al. “Microbial dark matter coming to light: challenges and opportunities”. In: 8.3 (Dec. 2020). DOI: 10.1093/nsr/nwaa280. URL: https://doi.org/10.1093/nsr/nwaa280.

[13] Dana L. Carper et al. “Cultivating the Bacterial Microbiota of Populus Roots”. In: (June 2021). Ed. by Stephen R. Lindemann. DOI: 10.1128/msystems.01306-20. URL: https://doi.org/10.1128/msystems.01306-20.

[14] Andrew D. Steen et al. “High proportions of bacteria and archaea across most biomes remain uncultured”. In: 13.12 (Aug. 2019), pp. 3126–3130. DOI: 10.1038/s41396-019-0484-y. URL: https://doi.org/10.1038/s41396-019-0484-y.

[15] Evan Bolyen et al. "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2". In: 37.8 (July 2019), pp. 852–857. DOI: 10.1038/s41587-019-0209-9. URL: https://doi.org/10.1038/s41587-019-0209-9.

[16] Angélique Rat et al. "Endophytic Bacteria From the Roots of the Medicinal Plant Alkanna tinctoria Tausch (Boraginaceae): Exploration of Plant Growth Promoting Properties and Potential Role in the Production of Plant Secondary Metabolites". In: 12 (Feb. 2021). DOI: 10.3389/fmicb.2021.633488. URL: https://doi.org/10.3389/fmicb.2021.633488.

[17] Martina Oberhofer et al. "Exploring Actinobacteria Associated With Rhizosphere and Endosphere of the Native Alpine Medicinal Plant Leontopodium nivale Subspecies alpinum". In: 10 (Nov. 2019). DOI: 10.3389/fmicb.2019.02531. URL: https://doi.org/10.3389/fmicb.2019.02531.

[18] Yupei Liu et al. "Engineering banana endosphere microbiome to improve Fusarium wilt resistance in banana". In: 7.1 (May 2019). DOI: 10.1186/s40168-019-0690-x. URL: https://doi.org/10.1186/s40168-019-0690-x.

[19] Nicolás D. Franco-Sierra et al. "Bacillus subtilis EA-CB0575 genome reveals clues for plant growth promotion and potential for sustainable agriculture". In: *Functional & Integrative Genomics* 20.4 (Mar. 2020), pp. 575–589. DOI: 10.1007/s10142-020-00736-x. URL: https://doi.org/10.1007/s10142-020-00736-x.

[20] Adarve-Rengifo Isabel. "Evaluation of endophytic microbiota of Calcutta 4 and Williams banana plants at in vitro and greenhouse level". MA thesis. Universidad EAFIT, Dec. 2020.

[21] Tobias Rausch et al. "Tracy: basecalling, alignment, assembly and deconvolution of sanger chromatogram trace files". In: 21.1 (Mar. 2020). DOI: 10.1186/s12864-020-6635-8. URL: https://doi.org/10.1186/s12864-020-6635-8.

[22] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: 17.1 (May 2011), p. 10. DOI: 10.14806/ej.17.1.200. URL: https://doi.org/10.14806/ej.17.1.200.

[23] Philip Ewels et al. "MultiQC: summarize analysis results for multiple tools and samples in a single report". In: 32.19 (June 2016), pp. 3047–3048. DOI: 10.1093/bioinformatics/btw354. URL: https://doi.org/10.1093/bioinformatics/btw354.

[24] Benjamin J Callahan et al. "DADA2: High-resolution sample inference from Illumina amplicon data". In: 13.7 (May 2016), pp. 581–583. DOI: 10.1038/nmeth.3869. URL: https://doi.org/10.1038/nmeth.3869.

[25] Michael S. Robeson et al. "RESCRIPt: Reproducible sequence taxonomy reference database management for the masses". In: (Oct. 2020). DOI: 10.1101/2020.10.05.326504. URL: https://doi.org/10.1101/2020.10.05.326504.

[26] Christian Quast et al. "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools". In: 41.D1 (Nov. 2012), pp. D590–D596. DOI: 10.1093/nar/gks1219. URL: https://doi.org/10.1093/nar/gks1219.

[27] Frank Oliver Glöckner et al. "25 years of serving the community with ribosomal RNA gene reference databases and tools". In: 261 (Nov. 2017), pp. 169–176. DOI: 10.1016/j.jbiotec.2017.06.1198. URL: https://doi.org/10.1016/j.jbiotec.2017.06.1198.

[28] Donovan H Parks et al. "A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life". In: 36.10 (Aug. 2018), pp. 996–1004. DOI: 10.1038/nbt.4229. URL: https://doi.org/10.1038/nbt.4229.

[29] T. Z. DeSantis et al. "Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB". In: 72.7 (July 2006), pp. 5069–5072. DOI: 10.1128/aem.03006-05. URL: https://doi.org/10.1128/aem.03006-05.

[30] Nuala A. O'Leary et al. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". In: 44.D1 (Nov. 2015), pp. D733–D745. DOI: 10.1093/nar/gkv1189. URL: https://doi.org/10.1093/nar/gkv1189.

[31] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *Journal of machine learning research* 12.Oct (2011), pp. 2825–2830.

[32] Torbjørn Rognes et al. "VSEARCH: a versatile open source tool for metagenomics". In: 4 (Oct. 2016), e2584. DOI: 10.7717/peerj.2584. URL: https://doi.org/10.7717/peerj.2584.

[33] Stephen F. Altschul et al. "Basic local alignment search tool". In: 215.3 (Oct. 1990), pp. 403–410. DOI: 10.1016/s0022-2836(05)80360-2. URL: https://doi.org/10.1016/s0022-2836(05)80360-2.

[34] Nicholas A. Bokulich et al. "Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin". In: *Microbiome* 6.1 (2018), p. 90. DOI: 10.1186/s40168-018-0470-z. URL: https://doi.org/10.1186/s40168-018-0470-z.

[35] Jeffrey J Werner et al. "Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys". In: 6.1 (June 2011), pp. 94–103. DOI: 10.1038/ismej.2011.82. URL: https://doi.org/10.1038/ismej.2011.82.

[36] Kazutaka Katoh and Daron M Standley. "MAFFT multiple sequence alignment software version 7: improvements in performance and usability". In: *Molecular biology and evolution* 30.4 (2013), pp. 772–780. DOI: 10.1093/molbev/mst010.

[37] Alexandros Stamatakis. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies". In: *Bioinformatics* 30.9 (2014), pp. 1312–1313. DOI: 10.1093/bioinformatics/btu033. URL: http://dx.doi.org/10.1093/bioinformatics/btu033.

[38] Sophie Weiss et al. "Normalization and microbial differential abundance strategies depend upon data characteristics". In: *Microbiome* 5.1 (2017), p. 27. ISSN: 2049-2618. DOI: 10.1186/s40168-017-0237-y.

[39] Manuel E. Lladser, Raúl Gouet, and Jens Reeder. "Extrapolation of Urn Models via Poissonization: Accurate Measurements of the Microbial Unknown". In: 6.6 (June 2011). Ed. by Dongxiao Zhu, e21105. DOI: 10.1371/journal.pone.0021105. URL: https://doi.org/10.1371/journal.pone.0021105.

[40] Sarah L. Hagerty et al. "An empirically derived method for measuring human gut microbiome alpha diversity: Demonstrated utility in predicting health-related outcomes among a human clinical sample". In: 15.3 (Mar. 2020). Ed. by Pratibha V. Nerurkar, e0229204. DOI: 10.1371/journal.pone.0229204. URL: https://doi.org/10.1371/journal.pone.0229204.

[41] William H Kruskal and W Allen Wallis. "Use of ranks in one-criterion variance analysis". In: *Journal of the American statistical Association* 47.260 (1952), pp. 583–621.

[42] Cameron Martino et al. "A Novel Sparse Compositional Technique Reveals Microbial Perturbations". In: 4.1 (Feb. 2019). Ed. by Josh D. Neufeld. DOI: 10.1128/msystems.00016-19. URL: https://doi.org/10.1128/msystems.00016-19.

[43] Marti J Anderson. “A new method for non-parametric multivariate analysis of variance”. In: *Austral ecology* 26.1 (2001), pp. 32–46. DOI: https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x.

[44] Sebastien Theil and Etienne Rifa. “rANOMALY: AmplicoN wOrkflow for Microbial community AnaLYsis”. In: 10 (Jan. 2021), p. 7. DOI: 10.12688/f1000research.27268.1. URL: https://doi.org/10.12688/f1000research.27268.1.

[45] Huang Lin and Shyamal Das Peddada. “Analysis of compositions of microbiomes with bias correction”. In: 11.1 (July 2020). DOI: 10.1038/s41467-020-17041-7. URL: https://doi.org/10.1038/s41467-020-17041-7.

[46] Zongzhuan Shen et al. “Deep 16S rRNA Pyrosequencing Reveals a Bacterial Community Associated with Banana Fusarium Wilt Disease Suppression Induced by Bio-Organic Fertilizer Application”. In: 9.5 (May 2014). Ed. by Gabriele Berg, e98420. DOI: 10.1371/journal.pone.0098420. URL: https://doi.org/10.1371/journal.pone.0098420.

[47] Michael C. Macey et al. “Impact of plants on the diversity and activity of methylotrophs in soil”. In: 8.1 (Mar. 2020). DOI: 10.1186/s40168-020-00801-4. URL: https://doi.org/10.1186/s40168-020-00801-4.

[48] Hiroyuki Iguchi, Hiroya Yurimoto, and Yasuyoshi Sakai. “Interactions of Methylotrophs with Plants and Other Heterotrophic Bacteria”. In: 3.2 (Apr. 2015), pp. 137–151. DOI: 10.3390/microorganisms3020137. URL: https://doi.org/10.3390/microorganisms3020137.

[49] Marina G. Kalyuzhnaya et al. “Methylotenera mobilis gen. nov., sp. nov., an obligately methylamine-utilizing bacterium within the family Methylophilaceae”. In: 56.12 (Dec. 2006), pp. 2819–2823. DOI: 10.1099/ijs.0.64191-0. URL: https://doi.org/10.1099/ijs.0.64191-0.

[50] Bram Vekeman et al. “New Methyloceanibacter diversity from North Sea sediments includes methanotroph containing solely the soluble methane monooxygenase”. In: 18.12 (Sept. 2016), pp. 4523–4536. DOI: 10.1111/1462-2920.13485. URL: https://doi.org/10.1111/1462-2920.13485.

[51] Min-Jung Kwak et al. “Genome Information of Methylobacterium oryzae, a Plant-Probiotic Methylotroph in the Phyllosphere”. In: *PLOS ONE* 9.9 (Sept. 2014), pp. 1–14. DOI: 10.1371/journal.pone.0106704. URL: https://doi.org/10.1371/journal.pone.0106704.

[52] John A. Fuerst and Evgeny Sagulenko. “Beyond the bacterium: planctomycetes challenge our concepts of microbial structure and function”. In: 9.6 (May 2011), pp. 403–413. DOI: 10.1038/nrmicro2578. URL: https://doi.org/10.1038/nrmicro2578.

[53] Manish Kumar et al. “Biodiversity of methylotrophic microbial communities and their potential role in mitigation of abiotic stresses in plants”. In: 74.3 (Jan. 2019), pp. 287–308. DOI: 10.2478/s11756-019-00190-6. URL: https://doi.org/10.2478/s11756-019-00190-6.

[54] Damien P. Devos and Naomi L. Ward. “Mind the PVCs”. In: 16.5 (Apr. 2014), pp. 1217–1221. DOI: 10.1111/1462-2920.12349. URL: https://doi.org/10.1111/1462-2920.12349.

[55] Karel Petrzik, Sára Brázdová, and Krzysztof Krawczyk. “Novel Viruses That Lyse Plant and Human Strains of Kosakonia cowanii”. In: 13.8 (July 2021), p. 1418. DOI: 10.3390/v13081418. URL: https://doi.org/10.3390/v13081418.

[56] Yuka Munakata et al. "Composition and functional comparison of vetiver root endophytic microbiota originating from different geographic locations that show antagonistic activity towards Fusarium graminearum". In: 243 (Feb. 2021), p. 126650. DOI: 10.1016/j.micres.2020.126650. URL: https://doi.org/10.1016/j.micres.2020.126650.

[57] Nurul Shamsinah Mohd Suhaimi et al. "Antagonistic effects of biofilm-forming bacterial strains co-inoculated with blood disease pathogenic strain, Ralstonia syzygii subspecies celebesensis in banana plants". In: 148.1 (Oct. 2016), pp. 13–26. DOI: 10.1007/s10658-016-1064-x. URL: https://doi.org/10.1007/s10658-016-1064-x.

[58] Satoshi Ishii et al. "Noviherbaspirillum denitrificans sp. nov., a denitrifying bacterium isolated from rice paddy soil and Noviherbaspirillum autotrophicum sp. nov., a denitrifying, facultatively autotrophic bacterium isolated from rice paddy soil and proposal to reclassify Herbaspirillum massiliense as Noviherbaspirillum massiliense comb. nov." In: 67.6 (June 2017), pp. 1841–1848. DOI: 10.1099/ijsem.0.001875. URL: https://doi.org/10.1099/ijsem.0.001875.

[59] Julian Taffner et al. "Exploring the Microbiota of East African Indigenous Leafy Greens for Plant Growth, Health, and Resilience". In: 11 (Nov. 2020). DOI: 10.3389/fmicb.2020.585690. URL: https://doi.org/10.3389/fmicb.2020.585690.

[60] Abdul Latif Khan et al. "Bacterial endophyte Sphingomonas sp. LK11 produces gibberellins and IAA and promotes tomato plant growth". In: 52.8 (July 2014), pp. 689–695. DOI: 10.1007/s12275-014-4002-7. URL: https://doi.org/10.1007/s12275-014-4002-7.

[61] Cecilia Taulé et al. "Endophytic colonization of sugarcane (Saccharum officinarum) by the novel diazotrophs Shinella sp. UYSO24 and Enterobacter sp. UYSO10". In: 403.1-2 (Feb. 2016), pp. 403–418. DOI: 10.1007/s11104-016-2813-5. URL: https://doi.org/10.1007/s11104-016-2813-5.

[62] Shipeng Lu et al. "Ecophysiology of Fe-Cycling Bacteria in Acidic Sediments". In: 76.24 (Dec. 2010), pp. 8174–8183. DOI: 10.1128/aem.01931-10. URL: https://doi.org/10.1128/aem.01931-10.

[63] Youry Pii et al. "The interaction between iron nutrition, plant species and soil type shapes the rhizosphere microbiome". In: 99 (Feb. 2016), pp. 39–48. DOI: 10.1016/j.plaphy.2015.12.002. URL: https://doi.org/10.1016/j.plaphy.2015.12.002.

[64] A. Vardanyan et al. "Effect of Cell Lysis (CLs) Products on Acidophilic Chemolithotrophic Microorganisms and the Role of Acidocella Species". In: 34.10 (May 2017), pp. 916–922. DOI: 10.1080/01490451.2017.1300203. URL: https://doi.org/10.1080/01490451.2017.1300203.

[65] Michael O. Eze et al. "Metagenomic insight into the plant growth-promoting potential of a diesel-degrading bacterial consortium for enhanced rhizoremediation application". In: (Mar. 2021). DOI: 10.1101/2021.03.26.437261. URL: https://doi.org/10.1101/2021.03.26.437261.

[66] Qianxin Deng et al. "Rhizosphere Microbial Communities Are Significantly Affected by Optimized Phosphorus Management in a Slope Farming System". In: 12 (Sept. 2021). DOI: 10.3389/fmicb.2021.739844. URL: https://doi.org/10.3389/fmicb.2021.739844.

[67] Xiaoning Gao et al. "Rhizosphere Bacterial Community Characteristics over Different Years of Sugarcane Ratooning in Consecutive Monoculture". In: 2019 (Nov. 2019), pp. 1–10. DOI: 10.1155/2019/4943150. URL: https://doi.org/10.1155/2019/4943150.

[68] Stephan Köstlbacher et al. "Pangenomics reveals alternative environmental lifestyles among chlamydiae". In: 12.1 (June 2021). DOI: 10.1038/s41467-021-24294-3. URL: https://doi.org/10.1038/s41467-021-24294-3.

[69] Astrid Collingro, Stephan Köstlbacher, and Matthias Horn. "Chlamydiae in the Environment". In: 28.11 (Nov. 2020), pp. 877–888. DOI: 10.1016/j.tim.2020.05.020. URL: https://doi.org/10.1016/j.tim.2020.05.020.

[70] Imen Nouioui et al. "Genome-Based Taxonomic Classification of the Phylum Actinobacteria". In: 9 (Aug. 2018). DOI: 10.3389/fmicb.2018.02007. URL: https://doi.org/10.3389/fmicb.2018.02007.

[71] Pious Thomas and Aparna Chandra Sekhar. "Cultivation Versus Molecular Analysis of Banana (Musa sp.) Shoot-Tip Tissue Reveals Enormous Diversity of Normally Uncultivable Endophytic Bacteria". In: 73.4 (Nov. 2016), pp. 885–899. DOI: 10.1007/s00248-016-0877-7. URL: https://doi.org/10.1007/s00248-016-0877-7.