Data in Brief

Re-examining the Rhizosphere: A Validation Study of 16S rRNA Metagenomics Analysis in Robusta Coffee (Coffea canephora L.) from the Central Highlands, Vietnam

Dien Ethan Mach

[1]Department of Bioinformatics, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD

Received on 12/18/2023; revised on 12/18/2023; accepted on 12/18/2023.

ABSTRACT

Metagenomics and 16S rRNA analysis are powerful tools for understanding the taxonomy and functional capabilities of microbial communities in environmental samples. The original study by Dr. Dinh Minh Tran analyzed rhizospheric soil samples of Coffea canephora L. from the Central Highlands of Vietnam, illuminating the taxonomy and functions of the Rhizosphere [1]. Our current study replicates the original computational methodology to validate these findings. By analyzing their 16S rRNA data using a metagenomics platform, we confirm the taxonomy and function of the microbial communities initially reported with some differences.

1 INTRODUCTION

Brazil and Vietnam are the two largest exporters of coffee globally while Vietnam is the largest exporter of Robusta coffee. Approximately 73% of the coffee from Vietnam is grown within the central highlands. Studying microbial diversity and function within the Rhizosphere from the Central Highlands of Vietnam can provide insight into sustainable cultivation of agriculture and the environmental ecology. Previous studies have focused on the microbial diversity of rhizospheric bacteria from Coffea canephora; however, the studies were limited to bacteria that could be grown in cultures [1]. Many microbes cannot be cultured which restricts the understanding of microbial diversity within such an environment [4]. As a result, an alternative, computational method for studying microbial communities can provide a more wholistic picture of ecological diversity.

Metagenomics allows for the study of microbial diversity without the use of culturing techniques and allows for high-throughput data analysis [3]. One popular method for identifying microbial populations within the community is through targeted sequencing and analysis of the 16s rRNA gene fragments within metagenomes. 16S rRNA genes contain 9 hypervariable regions (V1-V9) which can be used to distinguish bacteria. The V2, V3, and the V6 regions are typically used for “maximum discriminating power and maximum heterogeneity” in community profiling of bacterial groups [4]. In this study, all hypervariable regions (V1-9) were amplified and analyzed.

QIIME2 is an open-source bioinformatics platform wrapping many metagenomic tools within, allowing for microbiome data science and visualization [5]. Meanwhile, PiCrust2 is another bioinformatics tool used for predicting and functionally annotating 16S marker sequences of bacteria [6]. The original and current study used both QIIME2, Python, and PiCrust2 for the analysis.

2 METHODS

2.1 Data collection and processing

The soil sampling, DNA isolation, metagenomic sequencing, and library preparation methods are all described in detail in the original study [1]. The metagenomic data is publicly available and accessed through the NCBI SRA archive (SRR17644439). The data preprocessing included demultiplexing, trimming, and filtering for quality control via bcl2fastq, Trimmomatic (version 0.39), and Cutadapt (version 2.10) [1]. To download the pre-processed data, SRA tools and the fasterq-dump method was used.

2.2 Taxonomic Profiling

For the taxonomic analysis, QIIME2 (version 2020.8) was utilized within a Docker container. To import the FASTQ reads into QIIME2, a manifest csv file was manually generated in a text editor and input into the QIIME2 tools import method (type = SampleData[PairedEndSequencesWithQuality], input-format = PairedEndFastqManifestPhred33). Although the denoise-single method was originally used, the data was available as paired-end sequences. As a result, the denoise-paired method (p-trunc-len-f = 0, p-trunc-len-r = 0) was used instead to cluster and dereplicate reads into amplicon sequence variants (ASVs). After downloading the Silva 138 99% OTUs full-length sequences and taxonomy files, QIIME2’s feature-classifier classify-consensus-blast method was used to align the reads for community profiling. The resulting .qza file was exported to a .tsv file using QIIME2’s export tool. After, a quick bash script using sed was used to remove taxonomic prefixes from the .tsv file. The resulting .tsv file was imported back to a .qza file and visualized using the q2-Krona plugin. Using QIIME2’s krona collapse-and-plot method, a .qzv file was generated which can be viewed using QIIME2’s web-based viewer tool (https://view.qiime2.org/).

2.3 Functional Annotation

PiCrust2 (version 2.3.0-b) was used for functional annotation of the rhizospheric microbes based on MetaCyc databases. From the previously described taxonomic analysis, an ASV sequence and feature table .qza file were generated. Using the QIIME export tool, the files could be exported to .fasta and .biom files, respectively. After, the full PiCrust2 pipeline (picrust2_pipeline.py) was run using the .fasta and .biom files, outputting pathway abundance files.

Using the PiCrust2 add_descriptions.py script, the pathway abundance files were annotated with a combination custom mapping file. PiCrust2 offers the default MetaCyc mapping file; however, the file does not contain hierarchical levels as shown in the original study. As a result, top and second level mapping were downloaded [7]. Then, a Python script called custom_map_gen.py was written to combine the top and second level mapping files, creating a combination mapping file. Upon annotating the pathway abundances, another Python script (tsv_converter.py) is used to reformat the pathway abundances file for KronaTools. To visualize, the KronaTools ktImportText method could be used to convert the .tsv file into an .html file, resulting in a Krona plot.

3 RESULTS

3.1 Challenges in Reproduction

In Tran's original study, Table 1 shows the total amount of analyzed reads, classified reads, and unclassified reads as 256,462, 256,357, and 105, respectively [1]. However, when accessing the publicly available data through the SRA archive, the number of reads provided was 181,508. Meanwhile, the number of identified reads were 99.56%, resulting in 180,709 classified reads and 799 unclassified reads (Table 2) [8].

Reads	Count
Total analyzed reads	256,462
Classified reads	256,357
Unclassified reads	105

Table 1: Tran's summary of analyzed, classified, and unclassified reads in this study [1].

Reads	Count
Total analyzed reads	181,508
Classified reads	180,709
Unclassified reads	799

Table 2: Summary of analyzed, classified, and unclassified reads in the current study.

3.2 Taxonomy

Taxonomic analysis comparing the original and current study (Fig.1 and 2) show 28 vs. 24 phyla identified. In the original study, the top six most abundant phylum were Proteobacteria at 26.4%, followed by Actinobacteriota (19.83%), Acidobacteriota (15%), Gemmatimonadota (10.35%), Chloroflexi (9.24%), and Myxococcota (6.97%). Meanwhile, in the current study, the top six most abundant were Proteobacteria at 29%, Actinobacteriota (26%), Acidobacteriota (10%), Gemmatimonadota (10%), Myxococcota (9%), and Bacteroidota (8%). Of note, Chloroflexi was identified at 2%.

In Figure 1, 119 bacterial orders were detected while figure 2 shows 96 orders. The top five most abundant orders as shown in Figure 1 were Burkholderiales at 10.74%, followed by Gemmatimonadales (10.22%), Acidobacteriales (7.15%), Rhizobiales (3.91%), and Frankiales (3.04%). Meanwhile, in Figure 2, the most abundant order was Burkholderiales at 13%, followed by Gemmatimonadales (10%), Acidobacteriales (7%), Chitinophagales (6%), and Frankiales (5%).

156 bacterial families were detected in the original study and in Figure 1. Gemmatimonadaceae (10.22%) was the most abundant followed by Xanthobacteraceae (2.96%), Haliangiaceae (2.71%), Nitrosomonadaceae (2.56%), and Chitinophagaceae (2.44%). Meanwhile, Figure 2 shows 138 families. The most abundant is Gemmatimonadaceae at 10%, Chitinophagaceae and uncultured at 6%, and Haliangiaceae and Comamonadaceae at 4%. Lastly, 242 genera were identified from Tran's study while 197 were found in the current study.

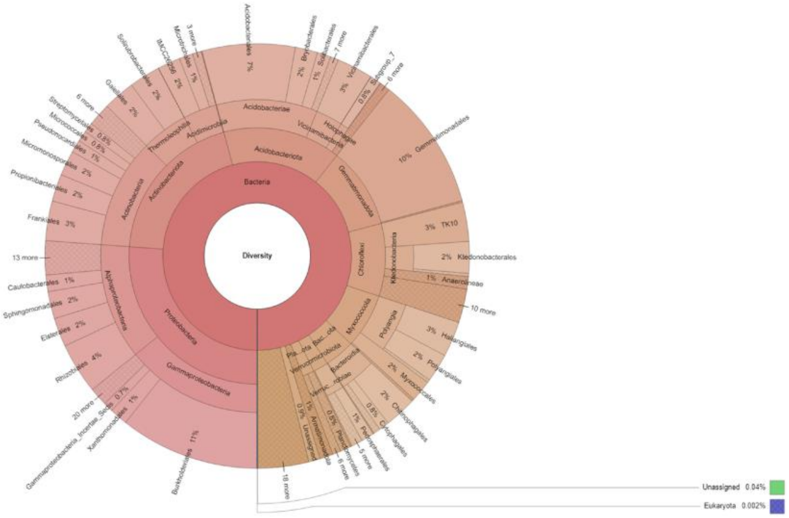

Figure 1: Tran's taxonomic profile of the rhizospheric microbiome of Coffea canephora L. in the Central Highlands region, Vietnam.

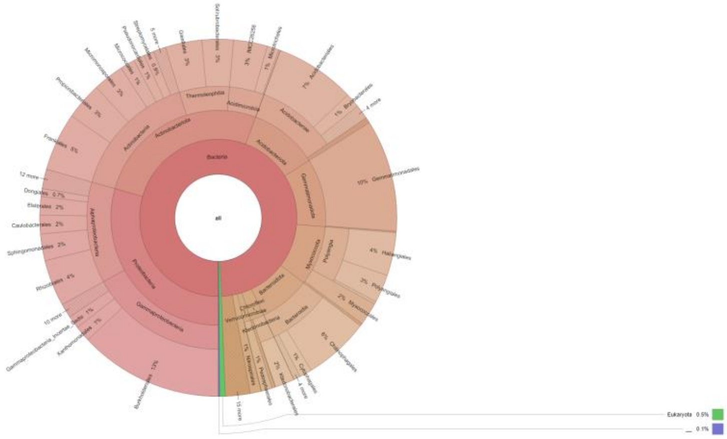

Figure 2: The current study's taxonomic profile of the rhizospheric microbiome of Coffea canephora L. in the Central Highlands region, Vietnam.

3.3 Functional Pathways

Originally, functional analysis of the original study (Figure 3) showed the primary function was biosynthesis at 69.67%, followed by degradation/utilization/assimilation (13.92%) and the generation of precursor metabolites and energy (12.92%). In the current study (Figure 4), the top three primary functions were biosynthesis at 69%, Degradation/Utilization/Assimilation at 15%, and Energy at 13%.

Within the functions related to biosynthesis shown in Figure 3, amino acid biosynthesis was highest at 16.91%, followed by cofactor, prosthetic group, electron carrier, and vitamin biosynthesis (15.51%), nucleoside and nucleotide biosynthesis (14.88%), fatty acid and lipid biosynthesis (8.8%), carbohydrate biosynthesis (5.02%), cell structure biosynthesis (3.43%), and secondary metabolite biosynthesis (2.75%). Meanwhile, in Figure 4, amino acid biosynthesis was highest at 17%, cofactor, prosthetic group, electron carrier, and vitamin biosynthesis was at 16%, nucleoside and nucleotide biosynthesis was at 14%, fatty acid and lipid biosynthesis was at 9%, carbohydrate biosynthesis was at 5%, cell structure biosynthesis was at 3%, and secondary metabolite biosynthesis was at 3%.

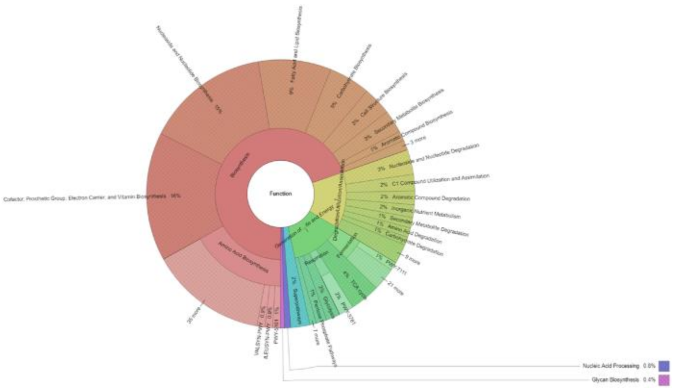

Figure 3: Original functional profiles of rhizosphere microbiome of Coffea canephora L. in the Central Highlands region, Vietnam.

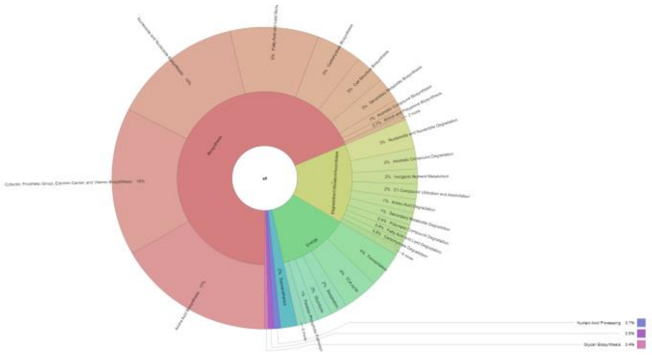

Figure 4: Current study's functional profiles of rhizospheric microbiome of Coffea canephora L. in the Central Highlands region, Vietnam.

4 DISCUSSION

This study aimed to replicate and validate the findings of Tran's original research on the rhizospheric microbiome of Coffea canephora L. in the Central Highlands region, Vietnam. Our findings largely align with the original study even though there are notable differences that warrant discussion.

4.1 Discrepancies in Read Counts

A discrepancy was observed in the total number of analyzed reads. While Tran's original study reported 256,462 reads, the current study found only 181,508. This variation is likely due to the pre-processing performed by the authors of the publicly available data.

4.2 Taxonomic Variations

As for the taxonomic analysis, there were slight differences between the abundance and diversity of bacterial phyla and orders. While the most abundant phyla and orders remained consistent, variations in their relative abundances were observed. These differences could be attributed to the differences in QIIME2 methodology performed. For example, the current study used the QIIME2 dada2 denoise-paired method instead of the denoise-single method. Moreover, for most of the QIIME2 pipeline analysis, default parameters were used since they were unspecified in the original paper. Additionally, reads were aligned against the Silva-138 99% OTU sequence and taxonomy files which may have differed from the original study's reference files.

4.3 Functional Pathway Analysis

Comparing the two functional analyses showed high congruence with biosynthesis, degradation/utilization/assimilation, and energy generation being the most predominant functions. Minor discrepancies in the percentages of these functions might be due to the parameters used in the PiCrust2 pipeline. Moreover, the hierarchical custom mapping files used in the current study were found on GitHub, leading to a difference in hierarchical levels shown in Figures 3 and 4 [7].

4.4 Implications for Agriculture and Ecology

Our study reinforces the complexity and diversity of the rhizospheric microbiome. Understanding these microbial communities is crucial for developing agricultural practices that promote plant health and yield, especially in a major coffee-producing region like the Central Highlands of Vietnam. Using this information, further studies can be performed focusing on soil health, nutrient cycling, disease suppression, plant growth, resilience, or even bioremediation.

4.5 Limitations of the Study

Although both studies performed 16S rRNA analysis to better understand the rhizospheric microbiome, the use of whole genome sequencing (WGS) analysis would have complemented the analyses. WGS analysis improves community profiling and can be more accurate than 16S rRNA analysis [9]. By analyzing both 16S rRNA and WGS data, a more wholistic understanding of the rhizospheric microbiome could be achieved. Moreover, the new methodology would aid in identifying rare bacterial species and genes [9].

Another limitation is the need for further studying of different soil samples. The original study used five soil samples, or 500 grams worth of soil. For a more representative understanding of the rhizosphere, additional sampling is required.

When comparing the original and current studies, the results were quite similar but with some discrepancies. A major reason for this is due to the way the methodology was described in the original paper. Many of the parameters used within the computational tools and custom files were not described in detail. As a result, when running the computational analysis for the current study, many methods were run using default parameters. This can pose a challenge to reproducibility in research, especially without specific domain knowledge. By providing further information about the methodology, the results could have been reproduced with higher accuracy.

Lastly, although the original study was written and published in 2022, the tools used could be updated. For example, the versions chosen for QIIME2 (v2020.8) and PiCrust2 (v2.3.0-b) were released in 2020. Using these versions presented challenges of their own as these versions are not as well supported as their most updated counterparts. Instead of using the standalone PiCrust2 version, using the most updated version of QIIME2 would have supported the use of the q2-PiCrust2 plugin which is specifically made for seamless use with the QIIME2 pipeline. The same can be said for using the q2-Krona plugin instead of the standalone KronaTools.

5 DATA AVAILABILITY

The metagenomic 16S rRNA data is publicly available through the NCBI SRA archive (SRR17644439). Additionally, the custom Python scripts and files required for the analysis is uploaded to a public GitHub repository (https://github.com/e10m/meta_final) [10].

6 REFERENCES

[1] Tran, D. M. (2022). Rhizosphere microbiome dataset of Robusta coffee (Coffea canephora L.) grown in the Central Highlands, Vietnam, based on 16S rRNA metagenomics analysis. Data in Brief, 42, 108106. https://doi.org/10.1016/j.dib.2022.108106

[2] Contributors to Wikimedia projects. (2012, June 20). List of countries by coffee production - Wikipedia. Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/List_of_countries_by_coffee_production

[3] Forbes, J. D., Knox, N. C., Ronholm, J., Pagotto, F., & Reimer, A. (2017). Metagenomics: the next culture-independent game changer. Frontiers in microbiology, 8, 1069.

[4] Shah N, Tang H, Doak TG, Ye Y (2011) Comparing bacterial communities inferred from 16s rRNA gene sequencing and shotgun metagenomics. Pac Symp Biocomput 165–176.

[5] Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., . . . Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology, 37(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

[6] Douglas, G. M., Maffei, V. J., Zaneveld, J. R., Yurgel, S. N., Brown, J. R., Taylor, C. M., Huttenhower, C., & Langille, M. G. I. (2020). PICRUSt2 for prediction of metagenome functions. Nature Biotechnology, 38(6), 685–688. https://doi.org/10.1038/s41587-020-0548-6

[7] GitHub - Jiung-Wen/picrust_mapping: mapping files for MetaCyc pathway. (n.d.). GitHub. https://github.com/Jiung-Wen/picrust_mapping

[8] SRA Archive: NCBI. (n.d.). SRA Archive: NCBI. https://trace.ncbi.nlm.nih.gov/Traces/index.html?view=run_browser&amp;page_size=10&amp;acc=SRR17644439&amp;display=analysis

[9] Ranjan, R., Rani, A., Metwally, A., McGee, H. S., & Perkins, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochemical and Biophysical Research Communications, 469(4), 967–977. https://doi.org/10.1016/j.bbrc.2015.12.083

[10] e10m/meta_final. (n.d.). GitHub. https://github.com/e10m/meta_final