

Original Research

Impact of Normalization Methods on Metagenomic Characterization of *Amaranthus Cruentus* var. Pribina-Associated Microbiomes Under Cadmium Stress

Dagmar Moravčíková ^{1,†}, Jana Žiarovská ^{1,†,*}, Alžbeta Žiarovská ²

1. Slovak University of Agriculture in Nitra, Faculty of Agrobiological Sciences, Institute of Plant and Environmental Sciences, Tr. A. Hlinku 2, 949 76, Nitra, Slovak Republic; E-Mails: xmoravcikova@uniag.sk; jana.ziarovska@uniag.sk
2. Slovak Technical University in Bratislava, Faculty of Informatics and Information Technologies, Ilkovičova 2, Karlova Ves, 842 16, Slovak Republic; E-Mail: alzbetaziarovska@gmail.sk

† These authors contributed equally to this work.

* **Correspondence:** Jana Žiarovská; E-Mail: jana.ziarovska@uniag.sk**Academic Editor:** Marilena Galdiero**Special Issue:** [Microbiome Analysis Techniques - Progress and Future](#)

OBM Genetics

2025, volume 9, issue 3

doi:10.21926/obm.genet.2503303

Received: January 17, 2025**Accepted:** July 04, 2025**Published:** July 15, 2025

Abstract

The study of endophytic and rhizosphere microbiota offers considerable potential for applications in agriculture, biotechnology, and bioremediation, given the phytoremediation capacity of *Amaranthus cruentus* var. Pribina performed a detailed analysis of the root and rhizosphere microbial communities under cadmium (Cd) stress. Although metagenomics provides powerful tools for microbial community profiling, the reproducibility and interpretability of the results are strongly influenced by the data processing strategies. In this study, special emphasis was placed on comparing normalization techniques and their effects on downstream analyses. Sequence data were processed in R using DADA2 to infer amplicon sequence variants (ASVs), followed by diversity, compositional, and statistical analyses using *phyloseq*, *vegan*, *ggplot2*, and *stats*. By evaluating multiple normalization approaches, it was



© 2025 by the author. This is an open access article distributed under the conditions of the [Creative Commons by Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is correctly cited.

demonstrated how the choice of method can significantly impact alpha and beta diversity metrics. These findings underscore the importance of normalization in microbiome studies and offer a tailored pipeline for analyzing Cd-induced shifts in plant-associated microbial communities.

Keywords

Metagenomics; endophytes; rhizosphere; Cd; R studio

1. Introduction

The rising global need for sustainable agriculture drives the search for innovative strategies that boost crop productivity while reducing environmental harm. One such promising strategy is the use of endophytes to improve plant growth, enhance nutrient absorption, and increase stress tolerance [1]. Metagenomics offers new opportunities to reveal the genetic diversity and function of microorganisms, enables accurate identification of non-culturable organisms, and helps to identify essential genes for ecological function [2]. Metagenomic studies have demonstrated that both environmental factors and the genotype of plants play roles in shaping the composition of the endophytic microbiome [3]. Given this complexity, selecting appropriate methods and procedures is essential, as the choice of procedures depends fundamentally on the research objective [4]. In metagenomics, two approaches are widely used: (A) sequencing only a specific region of 16S rRNA (Amplicon-based metagenomics) and (B) whole-genome sequencing (whole-genome metagenomics, also known as shotgun metagenomics). Each of these metagenomic strategies presents distinct advantages and limitations. The main advantages of amplicon-based metagenomics include that it's more affordable, the approach focuses on compartmentalization and diversity of microbial composition, and data processing is less challenging [5].

Nowadays, several tools are available for sequencing data processing, allowing users to simply upload their data to the server, where the metagenomes are automatically analyzed. The most well-known servers are MG-RAST [6] and GALAXY [7]. Additionally, there are web-based tools such as ANASTASIA [8] and SHAMAN [9], which are hosted on the GALAXY server. For researchers requiring more control over their analyses, several powerful command-line tools are available. QIIME 2 [10], which primarily runs on Unix-based systems (Linux and macOS) and is accessible on Windows through the Windows Subsystem for Linux (WSL), offers comprehensive microbiome analysis capabilities. Similarly, DADA2 [11], implemented as an R package, provides a flexible analysis environment compatible with all major operating systems. Python users can leverage the Sunbeam pipeline [12], which provides an alternative workflow for processing metagenomic data.

Researchers are increasingly interested in studying endophytes due to their potential applications in agriculture and biotechnology. However, traditional culture-dependent methods are severely limited, as they are capable of identifying only approximately 1% of environmental microbes. The use of omics technologies and metagenomics is crucial for uncovering the genetic potential of uncultured microbes, providing insights into their variability, structure, function, and metabolic pathways [13]. The successful analysis of microorganisms using next-generation sequencing relies on proper sampling procedures, the appropriate storage of root material, and the obtaining of DNA

of sufficient quality and quantity [14]. The study of endophytes using a metagenomics approach has already been investigated in several plant species such as *Eucalyptus urophylla* [15], *Fagopyrum esculentum* Moench [16], and *Panax ginseng* [17]. Simultaneously, when studying the composition of endophytes in plant roots, it is also beneficial to investigate the composition of the microbiome in the rhizosphere. By interacting with beneficial organisms, plants can optimize their growth and better withstand inhospitable conditions within the complex soil environment. Root secretions, which reflect a plant's adaptive strategies in response to its rhizosphere, enable plants to engage in positive or negative interactions, based on environmental factors, including the surrounding microbial communities [18]. The study of rhizosphere microbiome composition, as well as endophytes, has been investigated in various plants, including *Pinus massoniana* [19], *Salix* species [20], *Triticum aestivum*, *Hordeum vulgare* and *Secale cereale* [21]. The study of these plant-related microbiomes not only enhances understanding in the fields of plant and microbiology but also has practical application in agriculture [22] and phytoremediation techniques [23].

To date, no comprehensive studies have investigated the endophytic and rhizosphere microbial communities associated with *Amaranthus cruentus* var. *Pribina* under cadmium stress. This study is the first to specifically examine this cultivar, which holds promise in phytoremediation due to its ability to accumulate cadmium predominantly in the roots without visible toxicity symptoms. In addition to characterizing microbial changes induced by cadmium, this study provides a methodological contribution by systematically comparing three widely used normalization techniques for sequencing data. By combining a unique plant model with a comparison of data processing strategies, this research provides new insights into microbial interactions under heavy metal stress. It underscores the importance of methodological transparency in microbiome studies.

This study aims to evaluate the effect of three different normalization techniques on the assessment of alpha and beta diversity in microbial communities associated with *Amaranthus cruentus* var. *Pribina*. Specifically, the impact of median sequencing depth normalization, rarefaction, and percentage scale transformation was compared to determine which method best preserves biological signals while minimizing distortion.

2. Materials and Methods

2.1 Experimental Design

2.1.1 Plant Material

Amaranthus cruentus var. *Pribina* was used for the analysis. The growth of plants was conducted in a controlled chamber under the following conditions: a temperature of 23°C, 50% humidity, and a 16-hour light and 8-hour dark cycle. Root tissues were exclusively utilized for further analysis. Before experimenting, it is essential to plan the entire experimental design carefully. This includes considering the appropriate number of replicates, both biological and technical, to ensure a sufficient sample size for robust statistical analysis. Consistency in the preparation process is necessary to ensure the reproducibility and reliability of the research, as it involves following identical steps and procedures for all biological replicates [24]. In total, six pots were used. Each pot contained four plants, which were pooled into a single composite sample, resulting in six plant replicates. To accurately analyze endophytic communities, proper surface sterilization of root samples is critical. This step removes external contaminants, ensuring only true endophytes, those

living inside plant tissues, are studied [25]. Before DNA extraction, the roots were thoroughly rinsed with distilled water, surface-sterilized in a 2% sodium hypochlorite solution for 5 minutes, and then rinsed several more times with sterile distilled water to ensure complete sterilization.

2.1.2 Soil Material

To identify bacteria typical of the plant, it is necessary to analyze not only the plant itself but also the surrounding soil. In this study, rhizosphere samples were also included. The rhizosphere samples were collected from the plants as follows: plants were carefully removed from the pots, and the soil surrounding the roots of *Amaranthus* was carefully collected. The soil used for analysis was treated as if it were plants, with the pots kept in the growth chamber. The substrate used was a commercially available gardening soil (AGRO gardening substrate with active humus) and had the following characteristics: minimum burnable matter content. 70%, electrical conductivity in aqueous leachate 1w:25v max. 1.2 mS/cm, particle content above 20 mm max. 5.0%, moisture content max. 65%. To improve soil structure, sand was added in a 1:4 ratio. To improve soil structure, sand was added in a 1:4 ratio, and a 0.02 mol/dm³ CaCl₂ solution was added in a 1:2.5 ratio [26]. The resulting pH value was 5.443.

2.1.3 Cd Treatment

For the entire experiment, cadmium chloride (CdCl₂) was applied at a concentration of 15 mg/L. This concentration was chosen based on a previous study that tested various concentrations of CdCl₂ in *Amaranthus* spp. Including *Amaranthus cruentus* var. Pribina, which was also used in the study. The selected concentration did not induce any visible phytotoxic effects, and the majority of the cadmium was accumulated in root tissues [27]. Uniform experimental conditions, including the use of the same soil source, consistent biological replicates, and precise separation between control and treated groups, were maintained to ensure that observed differences in microbial diversity and composition could be confidently attributed to cadmium exposure rather than environmental variability.

2.1.4 DNA Extraction and Sequencing

Currently, numerous commercial DNA extraction kits are available, offering high-quality yields [3]. For DNA extraction, the NucleoSpin Soil kit (Nagel) was used. A quantity of 250–500 mg of soil or root material was weighed and transferred into a NucleoSpin® Bead Tube with ceramic beads, followed by the addition of 700 µL of Lysis Buffer. To enhance DNA yield, 150 µL of Enhancer SX was added, and the tube was vortexed vigorously for 5 minutes. After lysis, samples were centrifuged at 11,000 xg for 2 minutes. Following centrifugation, 150 µL of Buffer SL3 was added, and the samples were briefly chilled before undergoing another quick spin. Next, supernatant was filtered through a NucleoSpin® Inhibitor Removal Column at 11,000 xg. In the next step, 250 µL of Buffer SB was added, and the sample was loaded into a NucleoSpin® Soil Column. For purification, the column was sequentially rinsed with Buffer SB, SW1, and two rounds of SW2, spinning between each wash. After a final drying spin, the column was moved to a new tube, 100 µL of Buffer low TE was added, waited a minute, and the elution was completed with a spin, obtaining pure DNA, ready for analysis. The concentration of extracted genomic DNA was determined using a Qubit® fluorometer. Library

preparation and sequencing were provided by © SEQme s.r.o (Dobříš, Czech Republic) using Illumina. The primer pair 515F and 806R was used for amplification [28].

2.2 Bioinformatic Analysis

2.2.1 Data Preparation

Data preparation was performed in the RStudio environment. The DADA2 pipeline (version 1.30.0) (Divisive Amplicon Denoising Algorithm) [11] was used for sequencing data processing, including quality checking, trimming, merging paired-end reads, and removing chimeras. Initially, quality profiles of forward and reverse reads were inspected using the *plotQualityProfile* function, and a stringent Phred score cutoff of 30 was applied to retain only high-quality sequences. Following quality filtering and denoising, de novo chimera removal was performed, effectively eliminating artifactual chimeric sequences. Taxonomic assignment was conducted using the SILVA database [29]. In the context of DADA2, denoising refers to distinguishing sequencing errors from actual biological variation, thereby consolidating error-derived sequences into accurate amplicon sequence variants (ASVs) [30]. After processing the data with DADA2, a phyloseq object was constructed, containing metadata, DNASTrings, an ASV table, and a taxa table, using package *phyloseq* version 1.46.0. [31]. To further enhance data reliability and reduce potential contamination, taxa filtering was applied. A new object was created that contained taxa without the phylum *Cyanobacteria*. This was done using the *subset_taxa* command with a condition that excludes all taxa whose phylum-level classification corresponds to *Cyanobacteria*. The same approach was used to remove taxa assigned to the domains *Archaea* and *Eukaryota* at the kingdom level.

2.2.2 Data Normalisation

In microbiome studies, the number of species detected in samples is influenced by sequencing depth. Rare taxa may be represented with some samples simply because those samples were sequenced more thoroughly, generating a higher number of reads compared to others. These variations in total read numbers across samples prevent direct comparisons of absolute abundances. To address this, data normalization is essential before analysis. In this study, normalization was performed using the *transform_sample_counts* function from *vegan* package [32] in R. Normalization of count data involves dividing each entry into a count table by the total number of counts (sum of counts) for specific samples. This adjusts for differences in sequencing depth or other technical variations across samples. Additionally, various other normalization methods are commonly applied to microbial data in metagenomics [33]. In this study, specifically, read counts in each sample were normalized using the median sequencing depth, and rarefaction was conducted using the *rarefy_even_depth* function.

2.2.3 Data Visualization

For the visualization of Alpha diversity, the *ggplot* function from the package *ggplot2* (version 3.4.4) [34] was used. Function *geom_boxplot* was applied to create a boxplot for each normalization technique. For ordination analysis using NMDS, the *ordinate* function from the *phyloseq* package was used. Since ordination does not inherently produce graphical output, the *plot_ordination* function (*phyloseq*) was utilized to generate the respective visualizations.

2.2.4 Statistical Analysis

For statistical analysis, the stats package (version 4.3.1) [35] was used, specifically applying the *Kruskal-Wallis test*. *Test* function to perform the Kruskal-Wallis's rank sum test [36]. This non-parametric test was employed to compare alpha diversity results across different normalization techniques and to determine whether statistically significant differences existed between the methods.

3. Results

3.1 Alpha Diversity

Alpha diversity metrics, including indices of evenness, diversity and richness, are primarily base on the relative abundance of species within a sample. These metrics provide insight into the distribution and richness of species present in a particular sample [37]. To compare these indices within samples, it is essential to normalize data, as comparisons should be made between samples of equal size [38]. Among the most used indices are the Shannon and Simpson indices. The Shannon index was used to analyze heterogeneity in the samples (Figure 1), which is influenced by the rare species in the dataset [39], and the Simpson index, which depends on the common species [40].

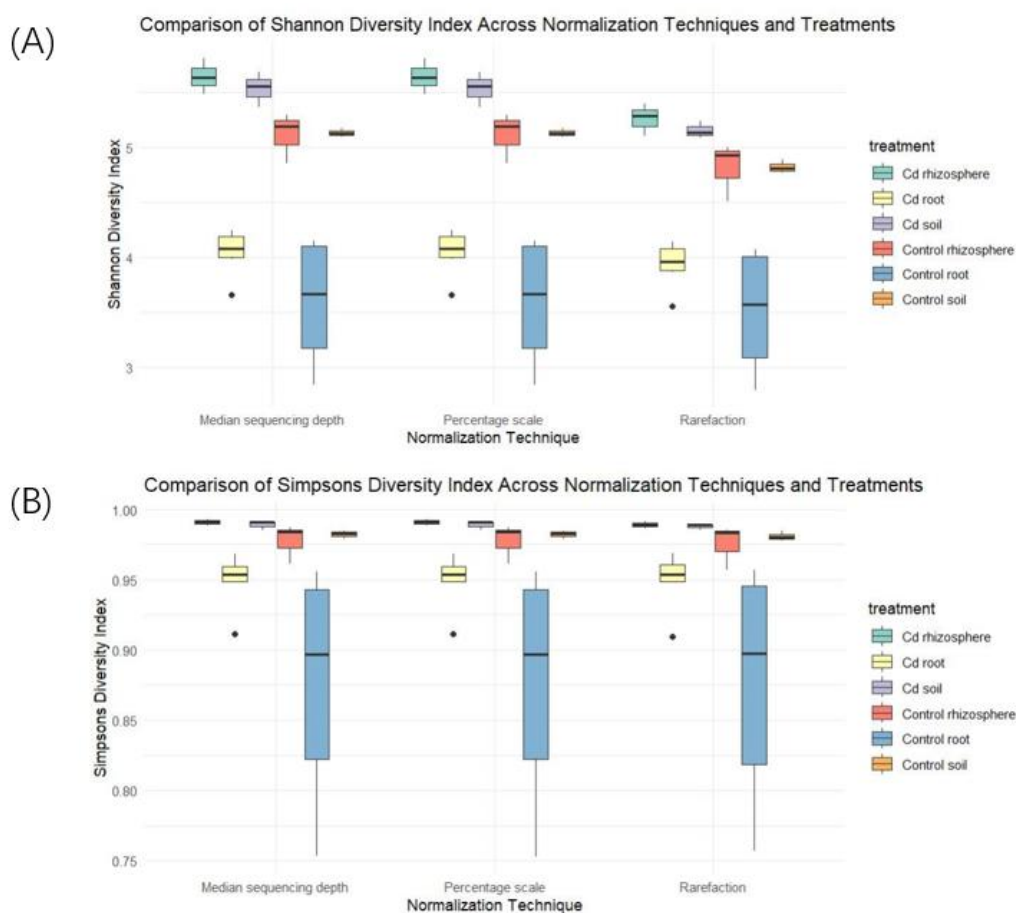


Figure 1 Comparison of Shannon diversity index across normalization techniques and treatment (A) and comparison of Simpson's diversity index across normalization

techniques and treatment (B). Boxplots show the median, quartiles, and potential outliers for each combination of method and treatment.

The comparison of alpha diversity across three different normalization techniques revealed only slight differences overall. The most notable differences were observed in the Shannon diversity index, where rarefaction led to a reduction in all treatments. In contrast, differences between median sequencing depth and the percentage scale method were minimal. Interestingly, the Simpson index displayed different behavior, indicating that the effect of normalization techniques on this index was negligible. To statistically assess the differences in alpha diversity between three normalization methods, the Kruskal-Wallis test was performed. The results indicated no statistically significant differences in alpha diversity, with p-values of 0.2984 for the Shannon index and 0.9895 for the Simpson index.

3.2 Beta Diversity

The NMDS algorithm operates through an iterative process, and for datasets that contain large amounts of data, using different starting values can occasionally lead to varying results [41]. Therefore, to ensure the reproducibility of the data, it is advisable to use *set.seed* function from the *base* package [35].

NMDS requires a distance matrix, and numerous dissimilarity coefficients have been developed to capture distinct aspects of variation between samples [42]. The decision regarding which coefficient to use for calculating distance matrices is crucial. In this type of research, the most suitable choice is often Bray-Curtis's dissimilarity calculation [43], which possesses many advantages. Cd treatment significantly altered the microbial composition in roots. At the genus level, a preliminary analysis suggests an increase in the relative abundance of *Streptomyces* and *Sphingopyxis* in Cd-treated roots (unpublished data).

This first graph (Figure 2A) illustrates the community structure following normalization based on median sequencing depth. The corresponding stress value is 0.056, indicating a relatively low level of distortion and suggesting that the relationships among microbial communities are visualized clearly and reliably. The second graph (Figure 2B) depicts the community structure following normalization to relative abundance. The associated stress value is 0.063, which is slightly higher than that of the median depth normalization. Nevertheless, the result still reflects a reasonable representation, although minor distortion may be present. The third graph (Figure 2C) shows the community structure of the dataset that was subjected to rarefaction. This approach resulted in a stress value of 0.082, the highest among the three methods. A higher stress value indicates more distortion in community structure, suggesting that rarefaction may not have fully captured the variability and complexity of the data.

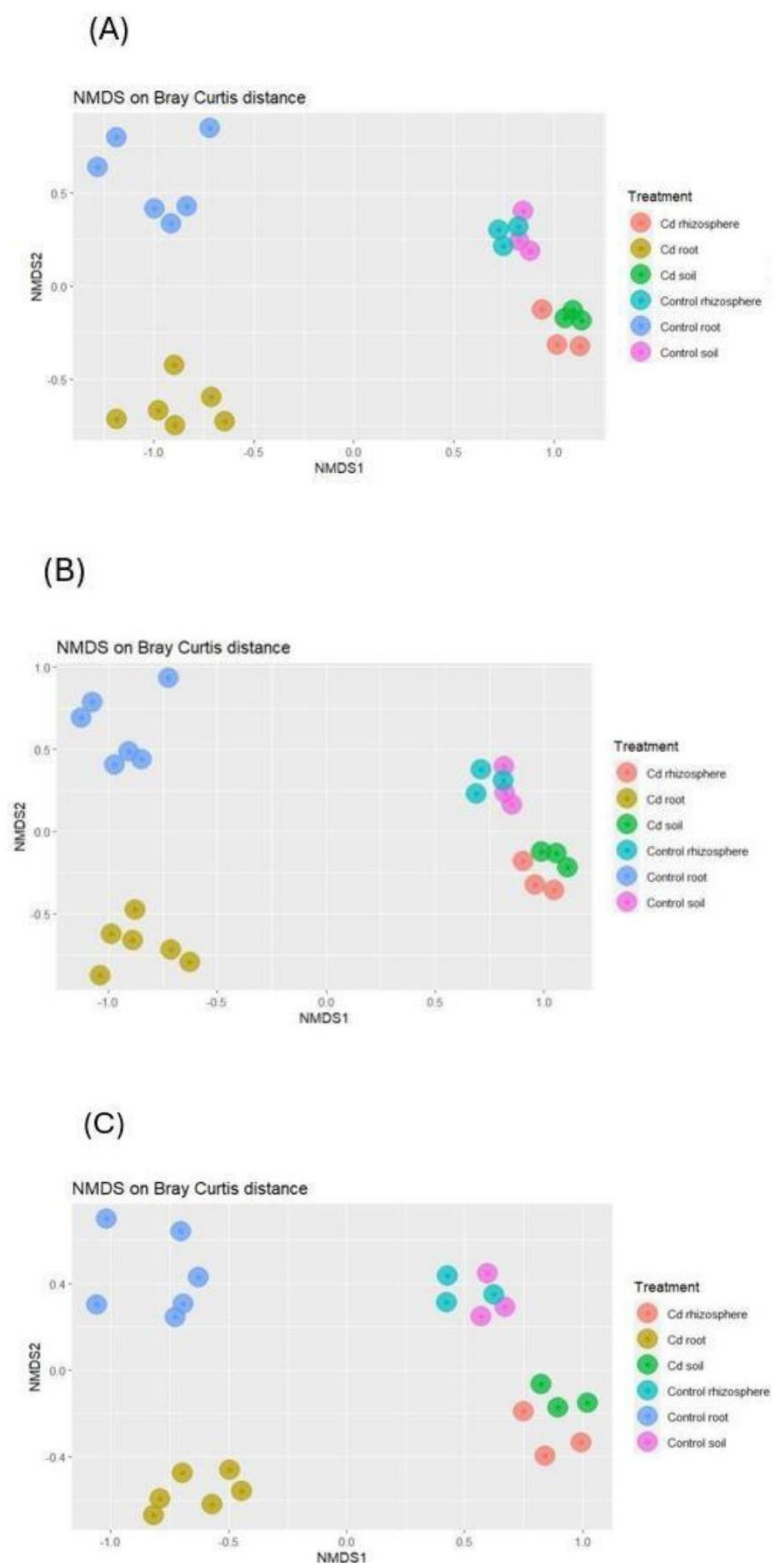


Figure 2 In the figure are NMDS graphs based on Bray-Curtis's dissimilarity matrix. The first graph represent of community structure after median sequencing depth (A) the stress = 0.056, the second one represent community structure after normalization in relative abundance (B) the stress = 0.063 and the last one represent the community structure of the dataset which was rarefied with the stress = 0.082.

4. Discussion

To ensure the robustness of our findings, we implemented a rigorous experimental design that included control samples and biological replicates. This allowed us to distinguish actual biological effects from methodological variability or data processing artifacts. Additionally, we applied and compared multiple normalization techniques to reduce biases introduced during preprocessing. The consistency of microbial shifts across both replicates and normalization methods strengthens the conclusion that the observed changes represent biologically meaningful responses to cadmium exposure. Consistent microbial shifts across replicates and across different normalization techniques support the conclusion that the observed changes represent accurate biological responses to Cd exposure.

Root surface sterilization was performed using a 2% sodium hypochlorite solution for 5 minutes. Alternative protocols, such as 70% ethanol for 3 minutes, have been proposed in previous studies. However, these require post-treatment verification to confirm the absence of residual microorganisms [44]. In the same survey, Tryptose Soya Agar (TSA) was used to check it. Two types of metagenomic analysis used to unravel the microbial identity and composition from high throughput sequencing data are: (i) amplicon-based analysis, which includes 16S ribosomal RNA for bacteria, internal transcribed spacer (ITS) and 18S region for fungi and eukaryotes, respectively, and (ii) whole metagenomic shotgun sequencing. A key advantage of whole-genome sequencing of endophytic microbes lies in its ability to reveal molecular and genetic traits that underlie their biological functions and determine preferences for colonizing specific environments [45]. In addition to the advantages of metagenomics, several limitations also exist when examining endophytic communities. One of the challenges is that when sequencing microbial DNA, plant DNA is often sequenced along with it. This creates a problem in distinguishing the microbial DNA sequences [13]. Once the sequence data have been processed and taxonomic classification assigned, it is essential to filter out potential contaminants. Failing to remove the sequence from eukaryotic organisms and cyanobacteria, mainly when focusing only on bacteria, can lead to false positives. For example, cyanobacterial sequences might be plant mitochondrial DNA, leading to misleading interpretations. Proper filtering is therefore essential to ensure the accuracy of the analysis [46]. Thus, *cyanobacteria*, eukaryotic organisms, and archaea were removed from the dataset, as the analysis focused exclusively on bacterial communities. However, some studies have introduced modified primers, making it unnecessary to remove cyanobacteria and eukaryotes during analysis. Locked oligonucleotide acid-PCR technique (LNA-PCR) has been used for this purpose, as well [47], and new plastid-specific LNA oligonucleotides have been proposed as being specific to plants. According to the study, these oligonucleotides appear to be very useful in investigating the endophytic microbiomes of plants [48]. It is essential to understand that these newly developed primers specifically inhibit the amplification of plastid DNA without adversely affecting the amplification of bacterial genes. The application of primers for mitochondrial and plastid sequences varies significantly. Oligonucleotides designed for mitochondrial sequences are applicable across a wide range of plant species (with a few exceptions). In contrast, oligonucleotides targeting plastid sequences are no longer universally applicable. Instead, these primers are categorized into three main groups based on the plant species for which they were designed [49]. Primers targeting the bacterial V4 variable region were used in the experiments, specifically the 515F and 806R primer pair. Several studies from other research fields have systematically compared different variable

regions to identify the most accurate region for bacterial detection [50]. Moreover, for the detection of endophytic bacteria, other studies have utilized different primer combinations. For instance, in *Passiflora incarnate*, V3-V4 region was targeted using an alternative set of primers [51]. Additionally, in Banana V3-V4 regions were targeted using primers which block the amplification of mitochondria and chloroplasts [52], and in *Pisum sativum* [44]. Due to the studies [53], a reliable overview of the basic primer combinations for endophyte analysis is available.

For the bioinformatic parts, there are lots of possibilities for how to analyze the metagenomics data. The SILVA database was used for sequence alignment in this study [54]. However, more databases exist that can be used in this type of research, such as the RDP classifier (Ribosomal Database Project Classifier) [55] and Greengenes [56]. For sequencing preparation, the DADA2 pipeline was used in this study. The concept of denoising involves distinguishing sequencing errors from genuine biological variations. By achieving this separation, denoising tools can consolidate sequences that only differ due to errors into a single, unified sequence. As a result, these tools produce single sequences referred to as amplicon sequence variants (ASVs), rather than reporting clusters of sequences. Compared to OTUs, ASVs facilitate direct sequence matching, making it more straightforward to compare data across different datasets. Nowadays, there is a growing preference in literature for using ASV over OTU [57, 58].

Three different normalization techniques were applied for data normalization: rarefaction, median sequencing depth, and proportional. Normalizing in relative abundance enables the inclusion of all taxa, regardless of their abundance, providing a comprehensive view of microbial composition. Pereira et al. [54] discussed several normalization techniques, including Trimmed Mean of M-values (TMM) and Relative Log Expression (RLE), which are widely used in the analysis of shotgun metagenomic data.

The two main components of analysis for metagenomic data are alpha and beta diversity. The impact of normalization methods was assessed using two commonly used alpha diversity indices: the Shannon and the Simpson index. Also typical is the Chao1 index, which considers singletons and doubletons in a sample. It uses the frequency of these rare species to estimate the total number of species present [59]. Thus, it is not possible to use this index after DADA2, because singletons are automatically removed in this pipeline [11]. The results from comparing the Simpson's index did not show any significant change between the different normalization techniques. The Simpson's index is a measure of biodiversity that primarily reflects the degree of dominance of particular species in a community. It emphasizes the concentration of abundance of one or a few species rather than a balanced representation of many species. This index is susceptible to the presence of dominant species, which can significantly affect the index value [60]. On the other hand, differences were observed in the Shannon index. These changes are the result of the background of the Shannon index. The natural logarithm of a small proportion is larger than the natural logarithm of a large proportion. Therefore, the Shannon index places additional importance on rare species by using the natural logarithm [61]. Rarefaction, as a normalization method based on random subsampling, often removes these rare taxa, especially in samples with low sequencing depth. As a result, Shannon diversity tends to decrease after rarefaction, potentially underestimating true community complexity and masking ecologically relevant variation. This observation was reflected in our results, where rarefied data consistently showed lower Shannon diversity values compared to other normalization techniques [62]. Moreover, its dependence on arbitrary library size thresholds and stochastic resampling adds unnecessary uncertainty. Therefore, its use should be carefully

considered, especially when alternative normalization approaches offer greater statistical reliability [63]. Median sequencing depth normalization involves scaling the read counts in each sample to the median total read count across all samples. This method retains all taxa and avoids random data loss, unlike rarefaction. Preserving low-abundance taxa and minimizing compositional distortion provides a more balanced view of microbial communities. Additionally, because it uses an actual reference point from the dataset (the median), it is less sensitive to outliers and extreme sequencing depths than proportional normalization. In our analysis, this method yielded the most consistent results in both alpha and beta diversity metrics, suggesting it effectively mitigated technical variability without compromising biological signal [64].

As already mentioned, beta diversity is also essential for understanding the differences between various samples. NMDS was employed for the visualization in this study. However, many studies [62] also utilize PCoA. The main difference between PCoA and NMDS is that PCoA uses eigenvalues to show how much variability is captured by each axis, which helps to assess the quality of the visualization. On the other hand, NMDS evaluates the quality of the result using the Stress values, which express the amount of distortion. A lower stress value means better preservation of the order of similarities between samples. PCoA is deterministic, meaning that repeated analyses on the same data will yield the same result, whereas NMDS is not [41]. The results can be evaluated based on the criteria described by Kruskal (1964) [65]: stress above 20% (>0.2): unfit - the model poorly represents the data. If stress is at 15% (>0.15) to 20% (>0.2) in this case, caution is required - representation is inadequate. Values of 10% (>0.1) to 15% (>0.15) are borderline - a better configuration would be desirable. Lower values of 5% (>0.05) to 10% (>0.10) are compliant, and the representation is acceptable. However, it is best if stress values are below 5% (<0.05). These values are classified as excellent because the model faithfully preserves the original distances. The observed differences in stress values highlight how each normalization technique addresses variability and community structure within the dataset. The median sequencing depth normalization produced the most reliable representation of the community structure, effectively reducing noise and preserving biological signals. The normalization of relative abundance introduced some variability that slightly distorted the community relationships. Rarefaction, while standardizing the total read counts, may have lost important information about rare taxa, resulting in greater distortion and a less accurate representation of community structure. These findings highlight the importance of selecting suitable normalization techniques in ecological and microbiome studies, as normalization methods can significantly impact the interpretation of community dynamics and relationships. Each normalization method used in the analysis of metagenomic data has its specific limitations that can affect the results and their interpretation. Although the rarefaction method standardizes sequencing depth across samples, it results in data loss, particularly for rare taxa, which can lead to an underestimation of biodiversity and increased variability in beta diversity. Most normalization methods enable the successful clustering of samples by biological origin when the groups differ significantly in their overall microbial composition. Rarefying cluster samples by biological origin more clearly than other normalization techniques for ordination metrics based on the presence or absence of taxa. Alternative normalization measures are potentially susceptible to artifacts due to library size, which is an essential consideration when interpreting results and selecting an appropriate method for a particular study [64]. Based on the findings of one publication, it has also been shown that rarefaction aligns better with unweighted distance metrics such as Jaccard or unweighted UniFrac. In contrast, scaling-based methods (e.g.,

proportional normalization) are more suited to weighted metrics but are more sensitive to library size differences and may introduce analytical artifacts [66].

The progress of the future research relies on expanding reference databases, platforms do meta-analysis and advancing bioinformatics tools capable of analyzing data from multiple perspectives [67]. Another advance in metagenomics is the implementation of machine learning. Machine learning offers notable advantages over traditional statistical approaches in microbial ecology. It is particularly effective at identifying subtle changes within microbial communities and isolating key bacterial taxa that are critical for predicting specific outcomes. Moreover, machine learning is capable of modeling complex and non-linear relationships between bacterial abundance data and environmental variables, mirroring the intricacy of real-world systems. This reduces the need for extensive data transformation or preprocessing, which is often a limiting factor in analyzing datasets [68].

5. Conclusions

The study focused on analyzing metagenomic data and investigating how this data behaves when three different normalization techniques are applied. Several functions in the R environment were used to assess their impact on the alpha and beta diversity of endophytes and microorganisms found in the roots of plants, specifically in *Amaranthus cruentus* var. *Pribina*. Several approaches currently exist for processing metagenomic data, enabling the detection of endophytes and microorganisms from both the rhizosphere and the soil. Each approach presents advantages and disadvantages, emphasizing the need for continuous development of bioinformatic methods for data analysis. The primary goal is to obtain accurate and reproducible results. Proper experimental design, the selection of suitable primers, the choice of isolation kits, and comprehensive sample preparation before sequencing are all equally important factors. These elements can significantly influence the quality and reliability of the resulting data. The selection of normalization techniques depends on the specific research question and characteristics of the dataset.

Author Contributions

D.M, J.Ž. – conceptualization, methodology, software, validation, formal analysis, D.M., A.Ž. – investigation, resources, writing original draft preparation, J.Ž. – writing review and editing, visualization; J.Ž. supervision and project administration

Funding

This study was supported by the project VEGA 2/0013/22 Amaranth plasticity in response to heavy metals: multi-scale analysis from ecophysiological to molecular aspects.

Competing Interests

The authors have declared that no competing interests exist.

Data Availability Statement

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

1. Das D, Sharma PL, Paul P, Baruah NR, Choudhury J, Begum T, et al. Harnessing endophytes: Innovative strategies for sustainable agricultural practices. *Discov Bact.* 2025; 2: 1.
2. Balan RP, Urumbil SK, Thomas S. Metagenomics: An advanced approach for endophyte study and prospecting. *Asian J Biol Life Sci.* 2024; 13: 599.
3. Olanrewaju OS, Glick BR, Babalola OO. Beyond correlation: Understanding the causal link between microbiome and plant health. *Heliyon.* 2024; 10: e40517.
4. Fadiji AE, Babalola OO. Metagenomics methods for the study of plant-associated microbial communities: A review. *J Microbiol Methods.* 2020; 170: 105860.
5. Rajguru B, Shri M, Bhatt VD. Exploring microbial diversity in the rhizosphere: A comprehensive review of metagenomic approaches and their applications. *3 Biotech.* 2024; 14: 224.
6. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, et al. The metagenomics RAST server—A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform.* 2008; 9: 386.
7. Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res.* 2020; 48: W395-W402.
8. Koutsandreas T, Ladoukakis E, Pilalis E, Zarafeta D, Kolis FN, Skretas G, et al. ANASTASIA: An automated metagenomic analysis pipeline for novel enzyme discovery exploiting next generation sequencing data. *Front Genet.* 2019; 10: 469.
9. Volant S, Lechat P, Woringer P, Motreff L, Campagne P, Malabat C, et al. SHAMAN: A user-friendly website for metataxonomic analysis from raw reads to statistical analysis. *BMC Bioinform.* 2020; 21: 345.
10. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol.* 2019; 37: 852-857.
11. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods.* 2016; 13: 581-583.
12. Clarke EL, Taylor LJ, Zhao C, Connell A, Lee JJ, Fett B, et al. Sunbeam: An extensible pipeline for analyzing metagenomic sequencing experiments. *Microbiome.* 2019; 7: 46.
13. Adeleke BS, Muller D, Babalola OO. A metagenomic lens into endosphere microbial communities, promises, and discoveries. *Lett Appl Microbiol.* 2023; 76: ovac030.
14. Romero-Salas EA, Ibarra-Sánchez CL. Isolation of Metagenomic DNA from Plant Roots. In: *Plant Microbiome Engineering*. New York, NY: Springer US; 2024. pp. 221-227.
15. de França Bettencourt GM, Degenhardt J, Dos Santos GD, Vicente VA, Soccol CR. Metagenomic analyses, isolation and characterization of endophytic bacteria associated with *Eucalyptus urophylla* BRS07-01 in vitro plants. *World J Microbiol Biotechnol.* 2021; 37: 164.
16. Li YL, He YJ, Li HM, Hu JJ, Cheng Z. Diversity of endophytes in *Fagopyrum esculentum* Moench. Seeds from different locations in China. *Russ J Plant Physiol.* 2021; 68: 413-420.

17. Hong CE, Kim JU, Lee JW, Bang KH, Jo IH. Metagenomic analysis of bacterial endophyte community structure and functions in *Panax ginseng* at different ages. 3 Biotech. 2019; 9: 300.
18. Walker TS, Bais HP, Grotewold E, Vivanco JM. Root exudation and rhizosphere biology. Plant Physiol. 2003; 132: 44-51.
19. Wu Y, Wang H, Peng L, Zhao H, Zhang Q, Tao Q, et al. Root-soil-microbiome interaction in the rhizosphere of Masson pine (*Pinus massoniana*) under different levels of heavy metal pollution. Ecotoxicol Environ Saf. 2024; 283: 116779.
20. Song X, Wang N, Zhou J, Tao J, He X, Guo N. High cadmium-accumulating *Salix* ecotype shapes rhizosphere microbiome to facilitate cadmium extraction. Environ Int. 2024; 190: 108904.
21. Lewin S, Wende S, Wehrhan M, Verch G, Ganugi P, Sommer M, et al. Cereals rhizosphere microbiome undergoes host selection of nitrogen cycle guilds correlated to crop productivity. Sci Total Environ. 2024; 911: 168794.
22. Xun W, Liu Y, Ma A, Yan H, Miao Y, Shao J, et al. Dissection of rhizosphere microbiome and exploiting strategies for sustainable agriculture. New Phytol. 2024; 242: 2401-2410.
23. Liu YQ, Chen Y, Li YY, Ding CY, Li BL, Han H, et al. Plant growth-promoting bacteria improve the Cd phytoremediation efficiency of soils contaminated with PE–Cd complex pollution by influencing the rhizosphere microbiome of sorghum. J Hazard Mater. 2024; 469: 134085.
24. Ju F, Zhang T. Experimental design and bioinformatics analysis for the application of metagenomics in environmental sciences and biotechnology. Environ Sci Technol. 2015; 49: 12628-12640.
25. Compant S, Cambon MC, Vacher C, Mitter B, Samad A, Sessitsch A. The plant endosphere world—bacterial life within plants. Environ Microbiol. 2021; 23: 1812-1829.
26. Rayment GE, Higginson FR. Australian laboratory handbook of soil and water chemical methods. Melbourne, Australian: Inkata Press; 1992. 330p.
27. Lancíková V, Tomka M, Žiarovská J, Gažo J, Hricová A. Morphological responses and gene expression of grain Amaranth (*Amaranthus* spp.) growing under Cd. Plants. 2020; 9: 572.
28. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci USA. 2011; 108: 4516-4522.
29. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 2012; 41: D590-D596.
30. Rosen MJ, Callahan BJ, Fisher DS, Holmes SP. Denoising PCR-amplified metagenome data. BMC Bioinform. 2012; 13: 283.
31. McMurdie PJ, Holmes S. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PloS One. 2013; 8: e61217.
32. Oksanen J, Simpson GL, Blanchet FG, Kindt R, Legendre P, Minchin PR, et al. Vegan: Community Ecology Package. R package Version 2.4-3 [Internet]. R Project; 2017. Available from: <https://CRAN.R-project.org/package=vegan>.
33. Wang B, Luan Y. Evaluation of normalization methods for predicting quantitative phenotypes in metagenomic data analysis. Front Genet. 2024; 15: 1369628.
34. Wickham H. ggplot2: Elegant graphics for data analysis. New York, NY: Springer; 2016.
35. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: The R Foundation; 2023. Available from: <https://www.R-project.org/>.

36. Hollander M, Wolfe DA. Nonparametric statistical methods. New York, NY: John Wiley & Sons; 1973.
37. Heip CH, Herman PM, Soetaert K. Indices of diversity and evenness. *Oceanis*. 1998; 24: 61-88.
38. Nolan KA, Callahan JE. Beachcomber biology: The Shannon Weiner Species Diversity Index. In: ABLE 2005 Proceeding. Claremont, CA: Association for Biology Laboratory Education; 2005. pp. 334-338.
39. Shannon CE. A mathematical theory of communication. *Bell Syst Technical J*. 1948; 27: 379-423, 623-656.
40. Simpson EH. Measurement of diversity. *Nature*. 1949; 163: 688.
41. Zuur AF, Ieno EN, Smith GM. Principal coordinate analysis and non-metric multidimensional scaling. In: *Analysing Ecological Data*. New York, NY: Springer; 2007. pp. 259-264.
42. Ricotta C, Podani J. On some properties of the Bray-Curtis dissimilarity and their ecological meaning. *Ecol Complex*. 2017; 31: 201-205.
43. Bray JR, Curtis JT. An ordination of the upland forest communities of southern Wisconsin. *Ecol Monogr*. 1957; 27: 326-349.
44. Hao J, Liu Q, Song F, Cui X, Liu L, Fu L, et al. Community Diversity of Endophytic Bacteria in the Leaves and Roots of Pea Seedlings. *Agronomy*. 2024; 14: 2030.
45. Kaul S, Sharma T, Dhar MK. “Omics” tools for better understanding the plant–Endophyte interactions. *Front Plant Sci*. 2016; 7: 955.
46. Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. *PLoS Comput Biol*. 2010; 6: e1000667.
47. Puri RR, Adachi F, Omichi M, Saeki Y, Yamamoto A, Hayashi S, et al. Metagenomic study of endophytic bacterial community of sweet potato (*Ipomoea batatas*) cultivated in different soil and climatic conditions. *World J Microbiol Biotechnol*. 2019; 35: 176.
48. Ikenaga M, Tabuchi M, Oyama T, Akagi I, Sakai M. Development of LNA oligonucleotide–PCR clamping technique in investigating the community structures of plant-associated bacteria. *Biosci Biotechnol Biochem*. 2015; 79: 1556-1566.
49. Ikenaga M, Sakai M. Application of locked nucleic acid (LNA) oligonucleotide–PCR clamping technique to selectively PCR amplify the SSU rRNA genes of bacteria in investigating the plant-associated community structures. *Microbes Environ*. 2014; 29: 286-295.
50. López-Aladid R, Fernández-Barat L, Alcaraz-Serrano V, Bueno-Freire L, Vázquez N, Pastor-Ibáñez R, et al. Determining the most accurate 16S rRNA hypervariable region for taxonomic identification from respiratory samples. *Sci Rep*. 2023; 13: 3974.
51. Goulart MC, Cueva-Yesquén LG, Hidalgo Martinez KJ, Attili-Angelis D, Fantinatti-Garboggini F. Comparison of specific endophytic bacterial communities in different developmental stages of *Passiflora incarnata* using culture-dependent and culture-independent analysis. *Microbiologyopen*. 2019; 8: e896.
52. Posada LF, Arteaga-Figueroa LA, Adarve-Rengifo I, Cadavid M, Zapata S, Álvarez JC. Endophytic microbial diversity associated with commercial cultivar and crop wild relative banana variety could provide clues for microbial community management. *Microbiol Res*. 2024; 287: 127862.
53. Tamošiūnė I, Andriūnaitė E, Stanys V, Baniulis D. Exploring diversity of bacterial endophyte communities using advanced sequencing technology. In: *Microbiome in Plant Health and Disease: Challenges and Opportunities*. Singapore: Springer; 2019. pp. 447-481.

54. Pereira MB, Wallroth M, Jonsson V, Kristiansson E. Comparison of normalization methods for the analysis of metagenomic gene abundance data. BMC Genom. 2018; 19: 274.
55. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007; 73: 5261-5267.
56. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006; 72: 5069-5072.
57. Fasolo A, Deb S, Stevanato P, Concheri G, Squartini A. ASV vs OTUs clustering: Effects on alpha, beta, and gamma diversities in microbiome metabarcoding studies. PloS One. 2024; 19: e0309065.
58. Jeske JT, Gallert C. Microbiome analysis via OTU and ASV-based pipelines—A comparative interpretation of ecological data in WWTP systems. Bioengineering. 2022; 9: 146.
59. Chao A. Nonparametric estimation of the number of classes in a population. Scand J Stat. 1984; 11: 265-270.
60. Whittaker RH. Evolution and measurement of species diversity. Taxon. 1972; 21: 213-251.
61. Chao A, Jost L. Coverage-based rarefaction and extrapolation: Standardizing samples by completeness rather than size. Ecology. 2012; 93: 2533-2547.
62. Zhang JZ, Li XZ, Yin YB, Luo SC, Wang DX, Zheng H, et al. High-throughput sequencing-based analysis of the composition and diversity of the endophyte community in roots of *Stellera chamaejasme*. Sci Rep. 2024; 14: 8607.
63. McMurdie PJ, Holmes S. Waste not, want not: Why rarefying microbiome data is inadmissible. PLoS Comput Biol. 2014; 10: e1003531.
64. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017; 5: 27.
65. Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika. 1964; 29: 1-27.
66. Pérez-Cobas AE, Gomez-Valero L, Buchrieser C. Metagenomic approaches in microbial ecology: An update on whole-genome and marker gene sequencing analyses. Microb Genom. 2020; 6: e000409.
67. De D, Nayak T, Das G, Dhal PK. Metagenomics and bioinformatics in microbial ecology: Current status and beyond. In: Applications of Metagenomics. Academic Press; 2024. pp. 359-385.
68. Kumar B, Lorusso E, Fosso B, Pesole G. A comprehensive overview of microbiome data in the light of machine learning applications: Categorization, accessibility, and future directions. Front Microbiol. 2024; 15: 1343572.