Table of Contents

[ASSIGNMENT | Application of Bioinformatics in Metagenomics 2](#_Toc188691374)

[1. Download FASTQ files, which correspond to the given twelve SRA IDs 3](#_Toc188691375)

[2. Check quality of the source data using FastQC and MultiQC 4](#_Toc188691376)

[3. Import FASTQ files into QIIME2 file format, pre-process and QC the data with QIIME2 tools 6](#_Toc188691377)

[4. Make ASVs feature table using DADA2 algorithm 7](#_Toc188691378)

[5. Evaluate phylogeny of identified microbial sequences 8](#_Toc188691379)

[6. Make rarefaction plots, select an appropriate rarefaction threshold 10](#_Toc188691380)

[7. Evaluate and interpret Alpha and Beta diversity in the dataset 11](#_Toc188691381)

[8. Evaluate the taxonomy of the soil microbial communities in the studied grassland sites 17](#_Toc188691382)

[9. Bibliography 20](#_Toc188691383)

# ASSIGNMENT | Application of Bioinformatics in Metagenomics

Matthew Spriggs

Applied Bioinformatics

ESTIMATED WORD COUNT (EXCLUDING FIGURES AND QUESTIONS): 991 words.

## Download FASTQ files, which correspond to the given twelve SRA IDs [marks: 3] [34 words]

### FASTQ files are downloaded using the scripts/s01\_get\_data\_assignment.sh script. This scrip will download FASTQ's from the NCBI Sequence Read Archive (SRA) database using the id's from data/samplesheet.csv as input. *Note: FASTQ's excluded from zip.*

### A diagram of a computer AI-generated content may be incorrect.

Figure 1: s01\_get\_data\_assignment.sh summary.

## Check quality of the source data using FastQC and MultiQC [marks: 3] [73 words]

### Following s01, quality checks are performed using FastQC tool and then MultiQC for aggregation (See data/\*.html files). Mean quality scores look good for all samples (Figure 2). Status checks look acceptable, with flags in Per Base Sequence Content, Per Sequence GC Content, Sequence Duplication Levels and Overrepresented sequences (Figure 3). These flags likely represent true characteristics for shotgun sequencing of 16s rRNA and do not indicate any  a problem with the data.

### A graph showing a number of scores AI-generated content may be incorrect.

Figure 2: Mean quality scores.

### A green and red bar chart AI-generated content may be incorrect.

Figure 3: FastQC Status Checks.

### A diagram of a file transfer AI-generated content may be incorrect.

Figure 4: s02\_qc.sh step summary.

## Import FASTQ files into QIIME2 file format, pre-process and QC the data with QIIME2 tools [marks: 3] [50 words]

### s03\_q2\_import\_and\_trim.sh is for processing metagenomic sequencing data using QIIME 2. It first preps data/source\_files\_local.txt, a samplesheet-type file that contains each sample-id from data/samplesheet.csv and the location of the forward and reverse primers. The script them imports paired-end sequence data, trims primers, and generates a visualization file for further analysis.

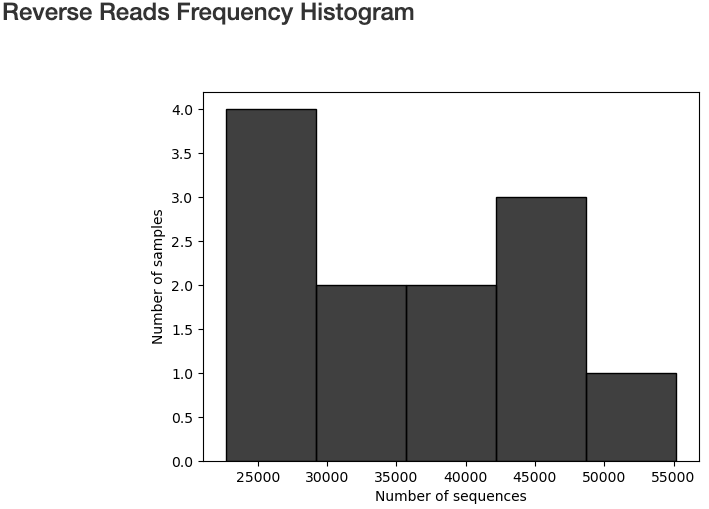
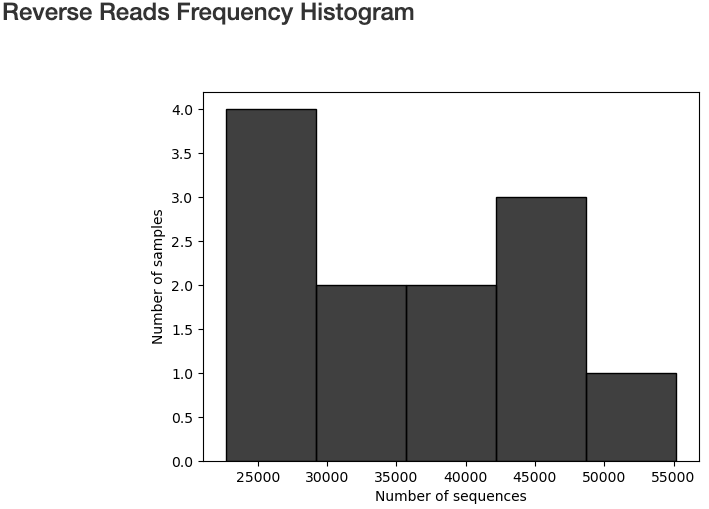
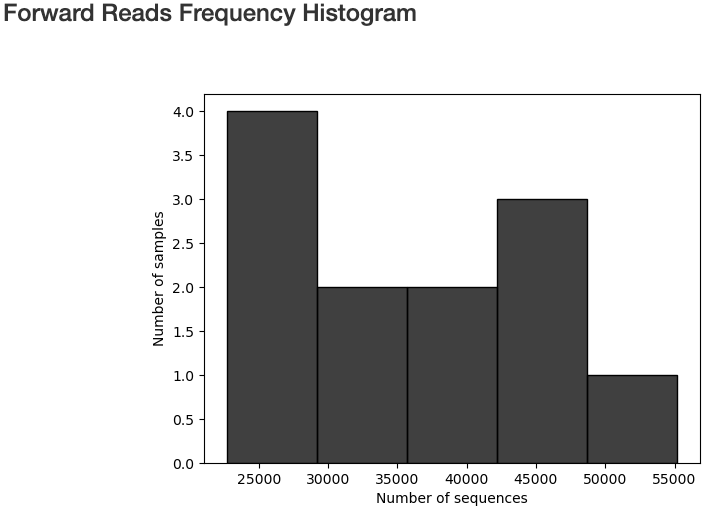
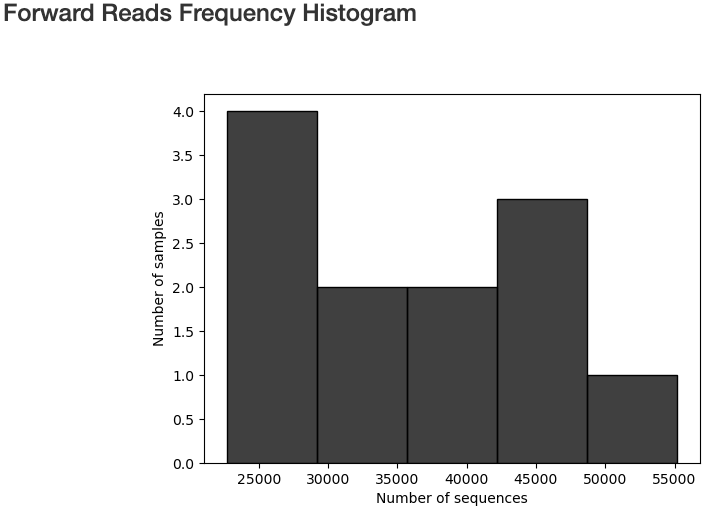


Figure 5: Read Frequence Histograms.

### A screenshot of a number AI-generated content may be incorrect.

Figure 6: Demultiplexed Sequence Counts Summary Table.

### A diagram of a computer AI-generated content may be incorrect.

Figure 7: s03\_q2\_import\_and\_trim.sh summary.

## Make ASVs feature table using DADA2 algorithm [marks: 3] [85 words]

### This script uses the DADA2 algorithm within QIIME 2 to clean and refine paired-end sequencing data. The goal is to remove noise and errors from the raw sequence data, resulting in high-quality, accurate sequences that can be used for downstream analysis, such as identifying and quantifying microbial species.

### Key outputs from this step are the features table (data/s04\_table\_dada2.qza) and metrics generated from data/ s04\_seqs\_dada2.qza. In this script, some of the max/min sequencing depth characteristics are exported into results/exported\_metadata/ sample-frequency-detail.csv for automation of s06.

### A screenshot of a computer AI-generated content may be incorrect.

Figure 8: s04\_q2\_denoise summary.

## Evaluate phylogeny of identified microbial sequences [marks: 3] [162 words]

### This step uses MAFFT alignment from QIIME2 to align sequences, combine with FastTree for phylogenetic tree construction. Masks alignment removes highly variable or uninformative regions. These regions can be problematic because they may not reflect true evolutionary relationships. (QIIME 2 development team, n.d.)

### Multiple sequence alignment arranges the sequences in a way that identifies regions of similarity, which can be indicative of functional, structural, or evolutionary relationships.

### FastTree algorithm constructs a phylogenetic tree from the masked aligned sequences, representing evolutionary relationships between the sequences. The aligned sequences and trees are then exported for visualization in external tools like the NCBI tree viewer See (supporting\_materials/s05\_tree\_nwk.pdf).

### The tree can be combined in R to make the dendrogram in Figure 9. To achieve this the user also needs the rarefied table output from step s06.

### The phylogenetic tree places gilb\_za and burrawan\_au samples into distinct cluster, indicating that the samples within these groups share closer evolutionary relationships with each other.

### A graph of numbers and a group of groups AI-generated content may be incorrect. Figure 9: Sample Dendrogram.

### A screenshot of a computer AI-generated content may be incorrect.

Figure 10: s05\_q2\_phylogenetic\_tree.sh step summary.

### A screenshot of a computer AI-generated content may be incorrect.

Figure 11: S11\_q2\_to\_R.

## Make rarefaction plots, select an appropriate rarefaction threshold [marks: 3] [109 words]

### Max-depth: Taken as max/min value for `*non-chimeric*` in supporting\_materials/metadata.tsv.

### Max-depth: `awk 'NR > 2 {print $8}' supporting\_materials/metadata.tsv | sort -r | head -n1`

### Min-depth: `awk 'NR > 2 {print $8}' supporting\_materials/metadata.tsv | sort | head -n1`

### See s06a\_alpha\_rarefaction.qzv for more interactive plots and other metrics.

### The threshold was calculated automatically to allow full automation of the pipeline.

### Max depth was chosen as the highest sequencing depth of non-chimeric reads from this file to ensure full coverage observed in the dataset.

### Sampling depth was chosen to match the minimal count of non-chimeric reads to ensure that all samples are included in the analysis.

### A screenshot of a computer AI-generated content may be incorrect.

Figure 12: s06a\_q2\_rarefaction\_plot.sh summary.

### A screenshot of a computer AI-generated content may be incorrect.

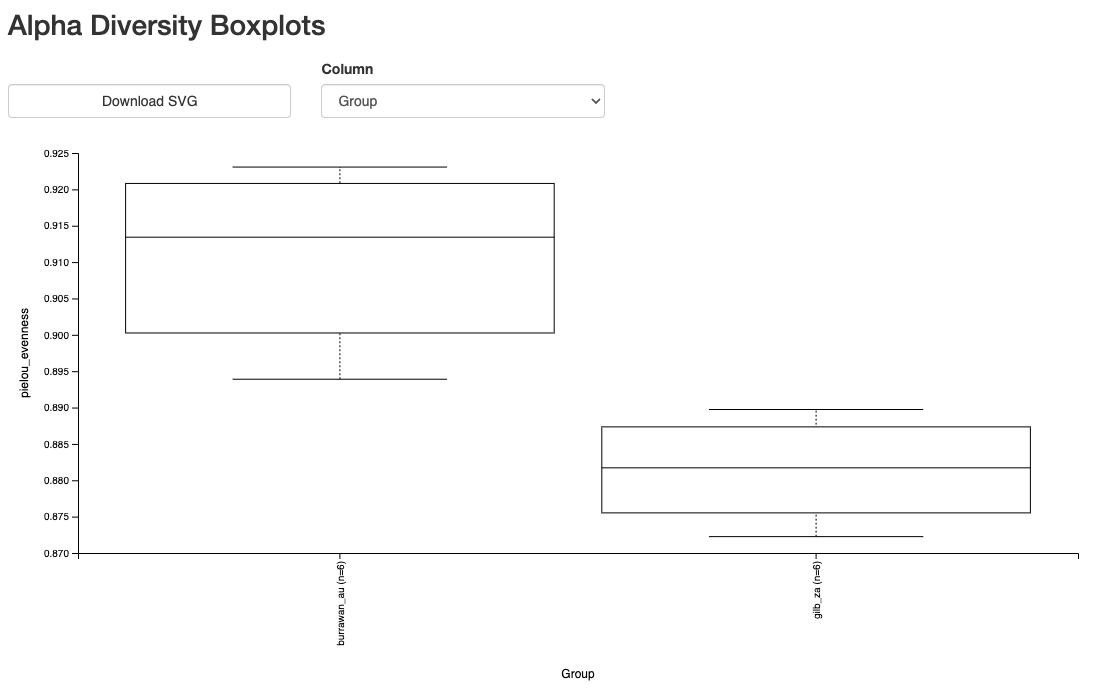
Figure 13: s06b\_q2\_apply\_rarefaction.sh summary.

## Evaluate and interpret Alpha and Beta diversity in the dataset [marks: 3] [284 words]

### Pielou\_eveness is statistically different between gilb\_za and burrawan\_au groups (0.004:1sf) for Kruskal-Wallis pairwise test (Figure 14).

### Pielou\_eveness is a metric used to help ecologists understand species evenness, e.g. how close in numbers different species are within an ecosystem. Values are distributed between 0-1, with 1 being all species of equal abundance (Chugani, 2025).

### burrawan\_au has a more even distribution of species than gilb\_za.



**s08\_alpha\_evenness\_per\_group.qzv**

**•**

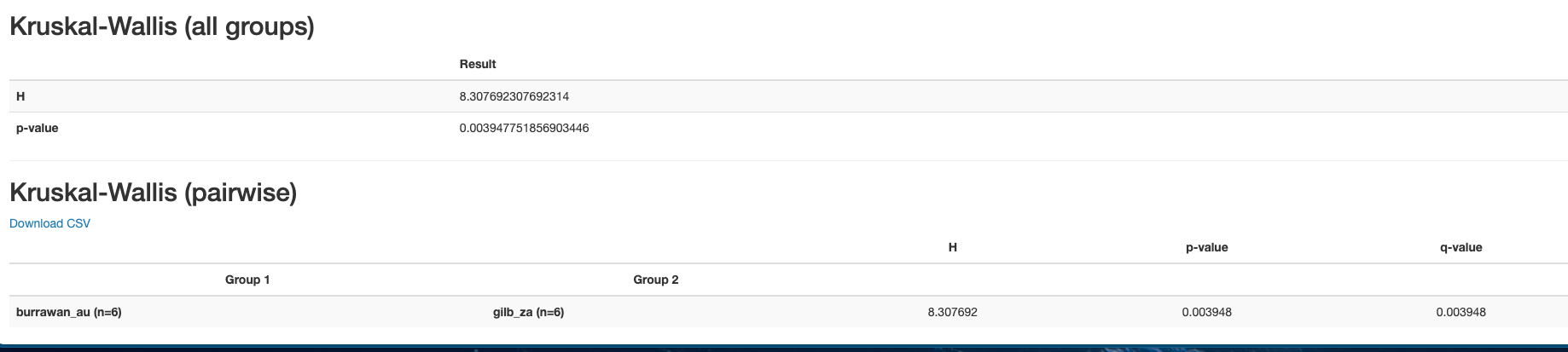
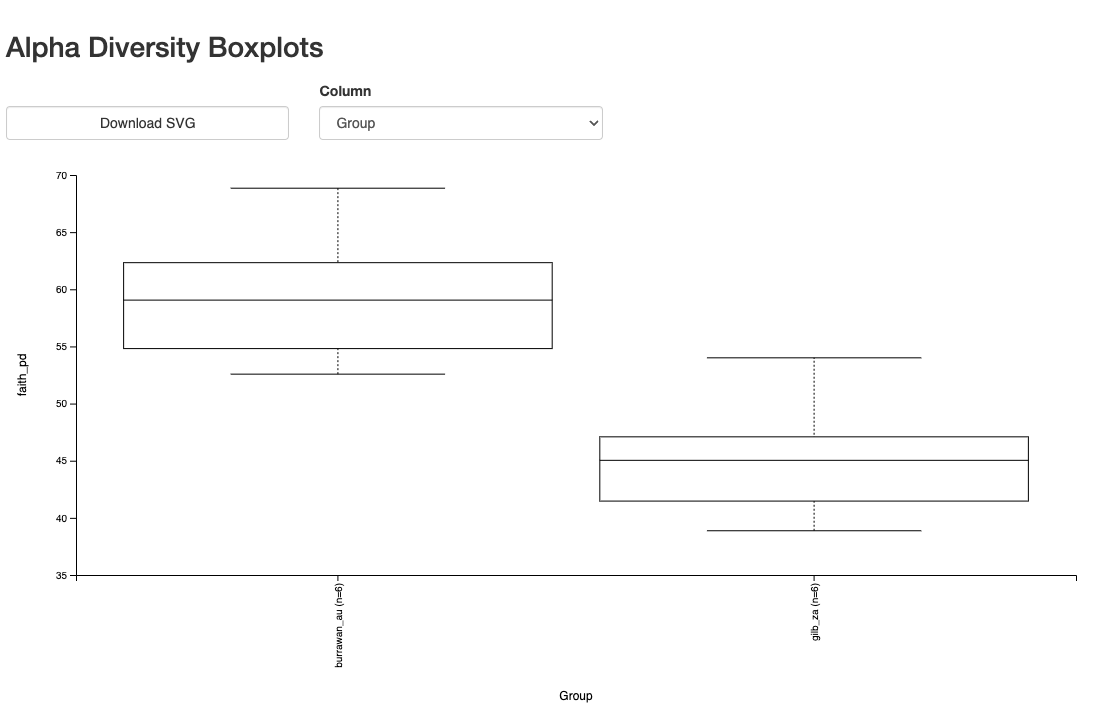


Figure 14: evenness\_vector.qza boxplots and significance tests between groups.

### Faith's Phylogenetic Diversity is a metric to assess the diversity of a community based on phylogeny. It represents a measure of branch lengths in a phylogenetic tree (Faith, 1992, as cited in Miller et al., 2018).

### Faith's Phylogenetic Diversity is statistically different between gilb\_za and burrawan\_au groups (0.006:1sf) for Kruskal-Wallis pairwise test (Figure 15).

### A higher value for this metric indicates greater branch lengths (inferred as greater phylogenetic diversity), meaning there is a greater breadth of bacterial community in the burrawan\_au group when compared to gilb\_za.



**s08\_alpha\_faith\_pd\_per\_group.qzv**

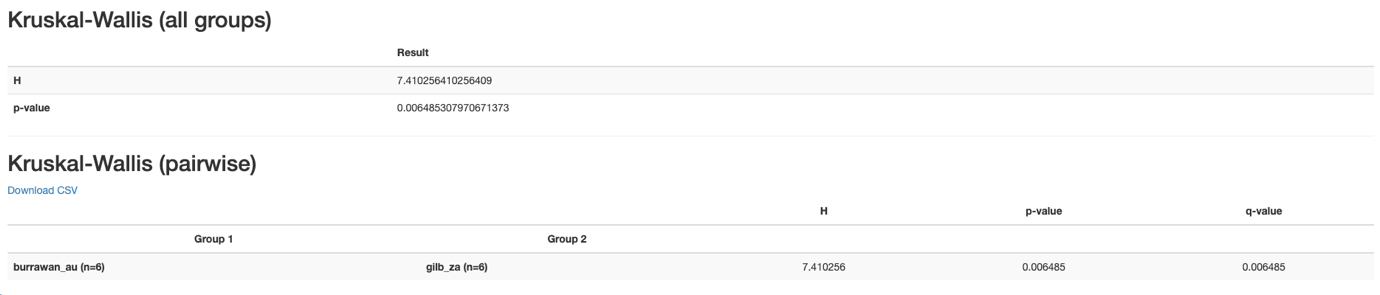
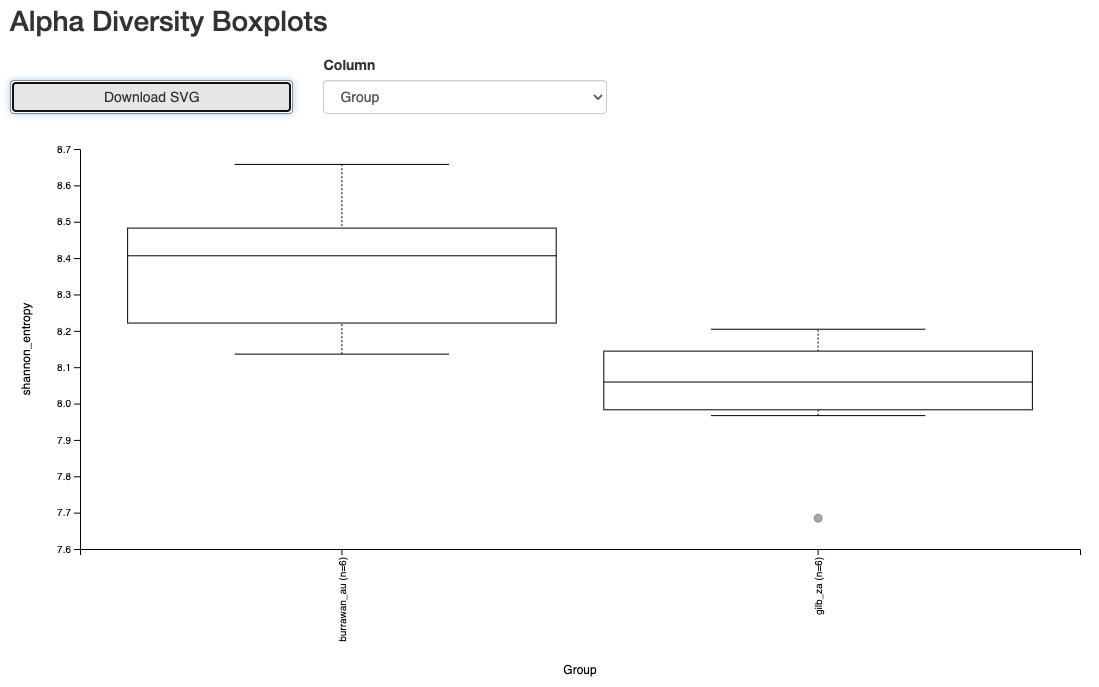


Figure 15: faith\_pd\_vector.qza boxplots and significance tests between groups.

### Shannon's Entropy is a measure of uncertainty or variability within a dataset (statisticshowto, n.d.). This metric is higher in burrawan\_au group when compared to gilb\_za (0.02:1sf) for Kruskal-Wallis pairwise test. Indicating greater diversity in burrawan\_au (Figure 16).



**s08\_alpha\_shannon\_per\_group.qzv**

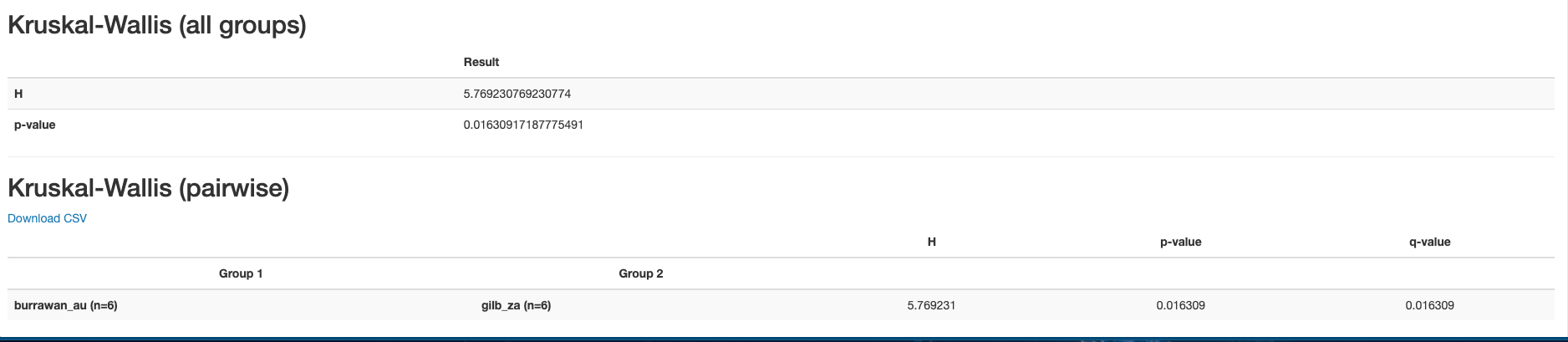
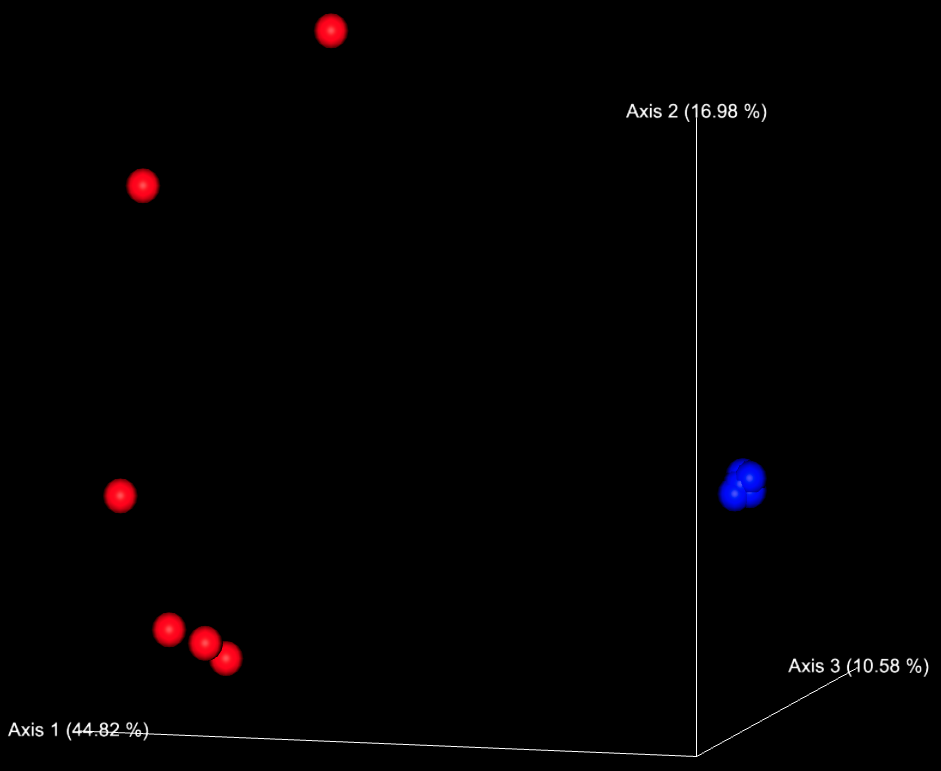


Figure 16: shannon\_vector.qza boxplots and significance tests between groups.

### Bray-Curtis Dissimilarity is a metric for species comparison between different communities (sites)(Wikipedia, n.d.). In the PCoA in Figure 17, samples from gilb\_za show more clustered diversity, indicating the microbial compositions of these samples is similar when compared to burrawan\_au samples.

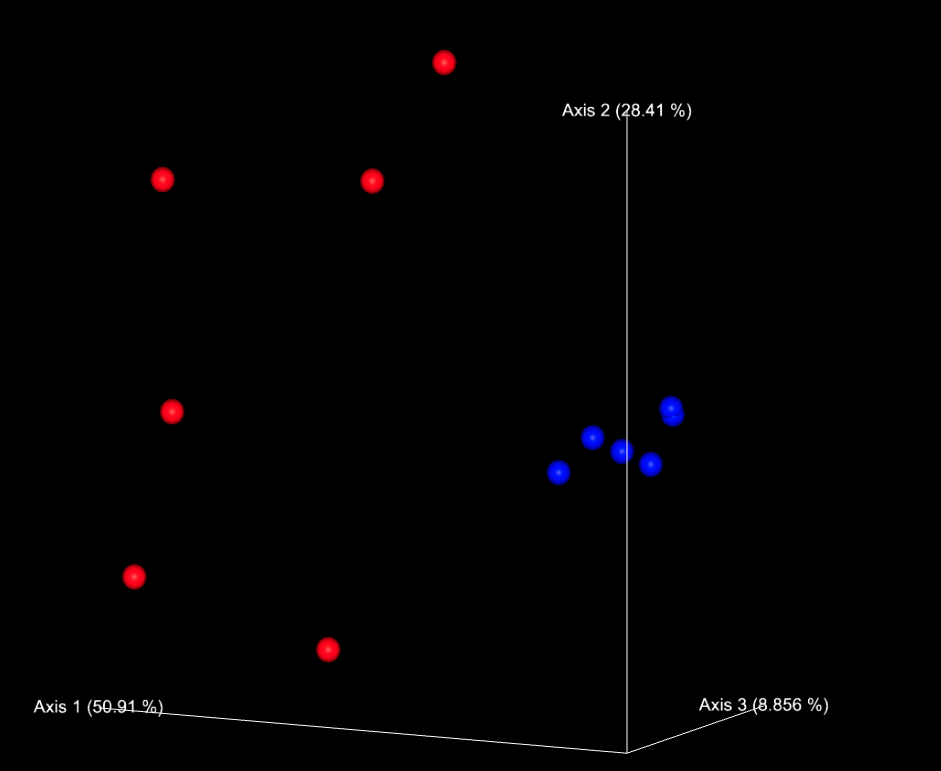


**s09\_beta\_bray\_curtis\_emperor\_pcoa.qzv**



Figure 17: PCoA from bray\_curtis\_pcoa\_results.qza.

### UniFrac is a beta-diversity metric that uses phylogenetic data to make comparisons between sites to understand relationship among microbial communities (Lozupone et al., 2011). Samples belonging to the gilb\_za cluster have similar species composition due to the tightly packed nature of this metric on the PCoA in Figure 18.



**s09\_beta\_weighted\_unifrac\_emperor\_pcoa.qzv**



Figure 18: PCoA from weighted\_unifrac\_emperor.qzv

### A screenshot of a computer AI-generated content may be incorrect.

Figure 19: s07\_q2\_calculate\_diversity\_metrics summary.

## Evaluate the taxonomy of the soil microbial communities in the studied grassland sites [marks: 3] [194 words]

### The file results/ s10\_taxa\_bar\_plot.qzv is an interactive plot that can be opened in QIIME2. For academic purposes the csv was downloaded from QIIME2 for level2 and loaded into R scripts/ metagenomics\_taxonomy\_plots.R to explore.

### Species diversity appears less in gilb\_za than burrawan\_au - although no quantitate analysis was performed (Figure 20).

### Several level-2 ASV's show high log fold changes between sites, perhaps indicative of different conditions (Figure 21). Although this is a poor metric to make comparisons, conclusions should be drawn from alpha-beta metrics on species composition differences.

### Drawing on conclusion made in (Leff et al., 2015), bacterial community composition is significantly influences by Nitrogen and Phosphorus additions, leading to an increase in *Actinobacteria, Alphaproteobacteria, and Gammaproteobacteria* and a decrease in *Acidobacteria, Planctomycetes, and Deltaproteobacteria.* The paper also notes that - although significant, bacterial diversity was not strongly altered by nutrient conditions (0.5% increase). The hypothesis was that increased nutrient availability will favour faster growing more efficient taxa and that these taxa would be more abundant. Next steps would include a deeper dive into the biology of the taxa from this report that show changes in alpha-beta metrics to confer results.

### 

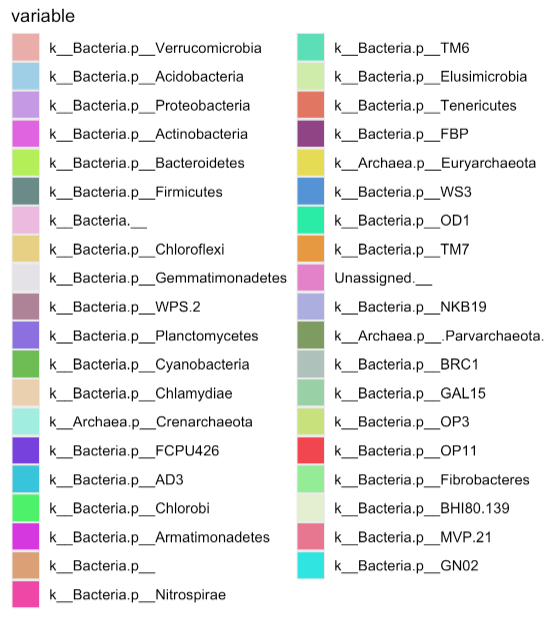
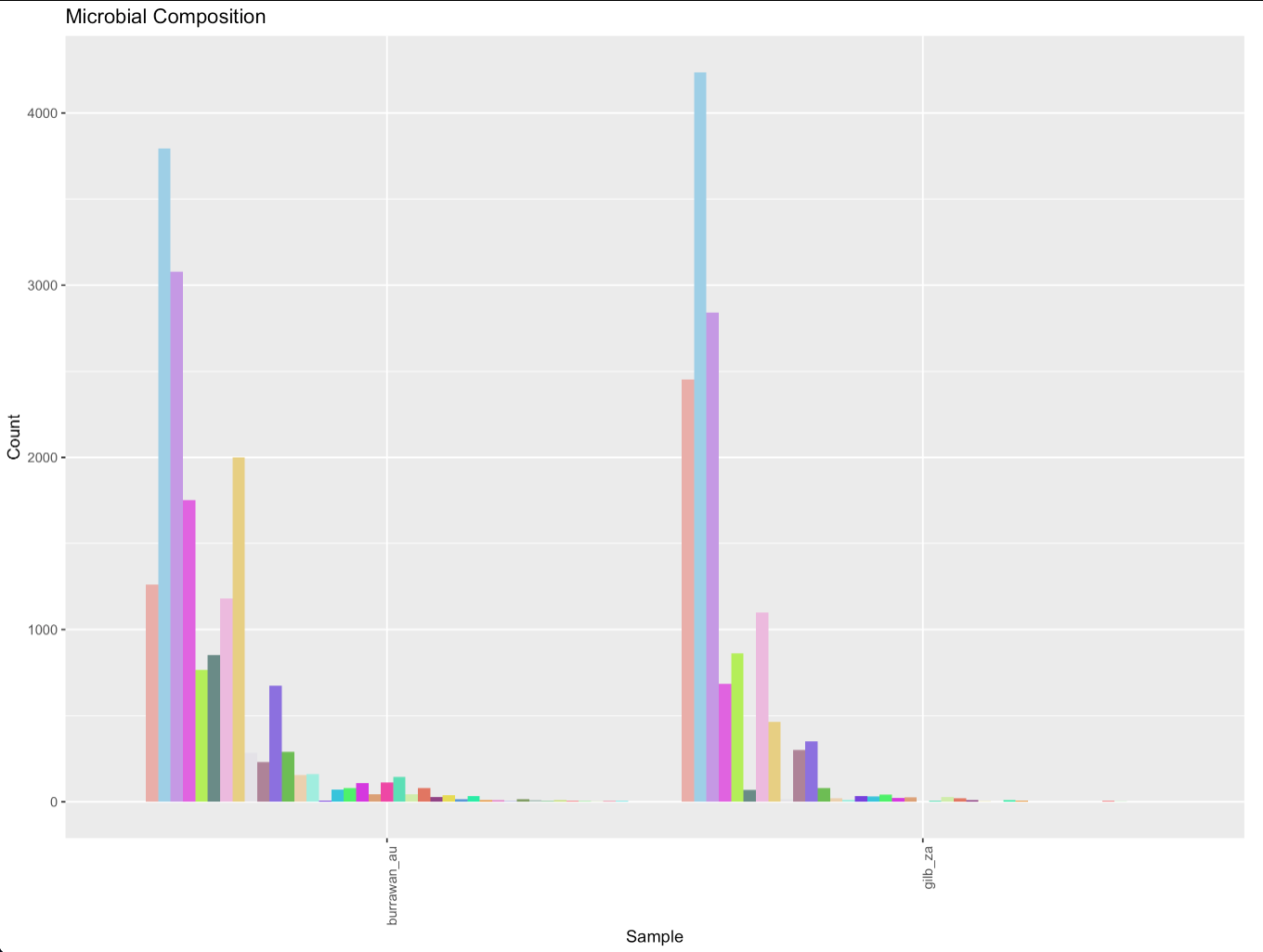


Figure 20: Level 2 Taxa barplot between groups.

### A graph with different colored lines AI-generated content may be incorrect.

Figure 21: Log change between sites.

## Bibliography

Chugani, V. (2025, January 6). *How to Calculate and Interpret Pielou’s Evenness Index*. https://www.statology.org/how-to-calculate-interpret-pielous-evenness-index/#:~:text=Pielou%E2%80%99s%20Evenness%20Index%20%28also%20known%20as%20the%20Shannon,in%20numbers%20different%20species%20are%20within%20an%20ecosystem.

Leff, J. W., Jones, S. E., Prober, S. M., Barberán, A., Borer, E. T., Firn, J. L., Harpole, W. S., Hobbie, S. E., Hofmockel, K. S., Knops, J. M. H., McCulley, R. L., La Pierre, K., Risch, A. C., Seabloom, E. W., Schütz, M., Steenbock, C., Stevens, C. J., & Fierer, N. (2015). Consistent responses of soil microbial communities to elevated nutrient inputs in grasslands across the globe. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(35), 10967–10972. https://doi.org/10.1073/pnas.1508382112

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., & Knight, R. (2011). UniFrac: An effective distance metric for microbial community comparison. In *ISME Journal* (Vol. 5, Issue 2, pp. 169–172). https://doi.org/10.1038/ismej.2010.133

Miller, J. T., Jolley-Rogers, G., Mishler, B. D., & Thornhill, A. H. (2018). Phylogenetic diversity is a better measure of biodiversity than taxon counting. *Journal of Systematics and Evolution*, *56*(6), 663–667. https://doi.org/10.1111/jse.12436

QIIME 2 development team. (n.d.). *QIIME2 Documentation*. Retrieved January 24, 2025, from https://docs.qiime2.org/2024.10/plugins/available/phylogeny/align-to-tree-mafft-fasttree/

statisticshowto. (n.d.). *Shannon Entropy*. Retrieved January 24, 2025, from https://www.statisticshowto.com/shannon-entropy/

Wikipedia. (n.d.). *Bray–Curtis dissimilarity*. Retrieved January 24, 2025, from https://en.wikipedia.org/wiki/Bray%E2%80%93Curtis\_dissimilarity#:~:text=In%20ecology%20and%20biology%2C%20the%20Bray%E2%80%93Curtis%20dissimilarity%20is,presented%20it%20in%20a%20paper%20in%201957.%20%5B1%5D