2)In another file (turn in a pdf) make a tutorial with screenshots explaining each of the input data and the results from the analyses. These should be done as figures with figure legends with 2-3 per page depending on the size. (If the figure or table is too big, just show a portion of it. In addition to the result images, also include figures and explanations for the following: a. The metadata file b. The barcode file c. A .qza file including citations and provenance d. The rep-seq file – see below on how to export e. demux.qvz – what is this and what does it show? f. stats-dada2 – what is dada2 and what does this mean? g. table.qvz – interactive sample detail (change sampling depth; feature detail)

-emp-single-end-sequences.qza

Metadata file :

sample-id	barcode-sequenc	body-site	year	month	day	subject	reported-antibiotic-usa	days-since-experiment-start
#q2:types	categorical	categorica	numeric	numeric	numeric	categorica	categorical	numeric
L1S8	AGCTGACTAGTC	gut	2008	10	28	subject-1	Yes	0
L1S57	ACACACTATGGC	gut	2009	1	20	subject-1	No	84
L1S76	ACTACGTGTGGT	gut	2009	2	17	subject-1	No	112
L1S105	AGTGCGATGCG	gut	2009	3	17	subject-1	No	140
L2S155	ACGATGCGACC/	left palm	2009	1	20	subject-1	No	84
L2S175	AGCTATCCACGA	left palm	2009	2	17	subject-1	No	112
L2S204	ATGCAGCTCAGT	left palm	2009	3	17	subject-1	No	140
L2S222	CACGTGACATGT	left palm	2009	4	14	subject-1	No	168
L3S242	ACAGTTGCGCG/	right palm	2008	10	28	subject-1	Yes	0
L3S294	CACGACAGGCT/	right palm	2009	1	20	subject-1	No	84
L3S313	AGTGTCACGGTC	right palm	2009	2	17	subject-1	No	112
L3S341	CAAGTGAGAGA(	right palm	2009	3	17	subject-1	No	140
L3S360	CATCGTATCAAC	right palm	2009	4	14	subject-1	No	168
L5S104	CAGTGTCAGGA(	tongue	2008	10	28	subject-1	Yes	0

Figure 0: his metadata file provides essential information about each sample, including sample ID, barcode sequence, and associated metadata such as body site, date of collection, and subject ID.The reported-antibiotic-usage column indicates whether the subject used antibiotics at the time of sampling, and days-since-experiment-start helps track the temporal aspect of sampling. These details are used throughout QIIME 2 analyses to group and compare samples based on factors like body site and antibiotic use.

Input data: qiime tools import \ → visualization file that provides an interactive feature table

--type EMPSingleEndSequences \ → mporting the sequence data

--input-path emp-single-end-sequences \ → import raw data

--output-path emp-single-end-sequences.qza → Specifies the file name

Output: emp-single-end-sequences.qza

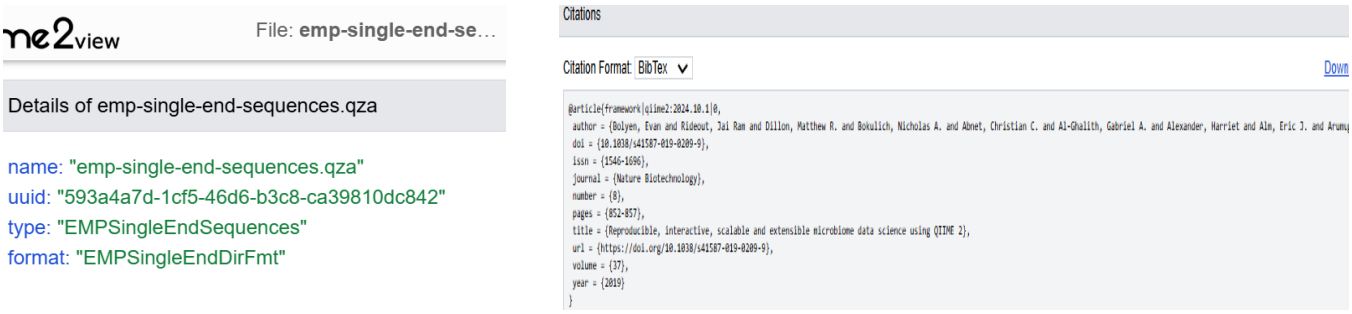


Figure 1: This is to ensure that the data is accurate. Name: is the name of the file. Uuid is the unique ID for the file. Type is to conform what the data is, and in this case is single sequencing data. The format is about the structure of Qiime. single-end sequencing data following the Earth Microbiome Project format, including both sequence and barcode reads. It serves as the initial input for demultiplexing, allowing assignment of sequences to their respective samples.

Demultiplexing sequences: To know about which barcode sequence is associated with each sample.

Input code is : qiime demux emp-single \

--i-seqs emp-single-end-sequences.qza \ ➔ Input the multiplexed sequences

--m-barcodes-file sample-metadata.tsv \ ➔ The metadata file (in TSV format)

--m-barcodes-column barcode-sequence \ ➔

Column name in metadata that contains each sample's unique barcode

--o-per-sample-sequences demux.qza \

--o-error-correction-details demux-details.qza

Output: Two files:

demux.qza➔ contain the demultiplexed sequences

demux-details.qza➔ This one is showing error correction details.

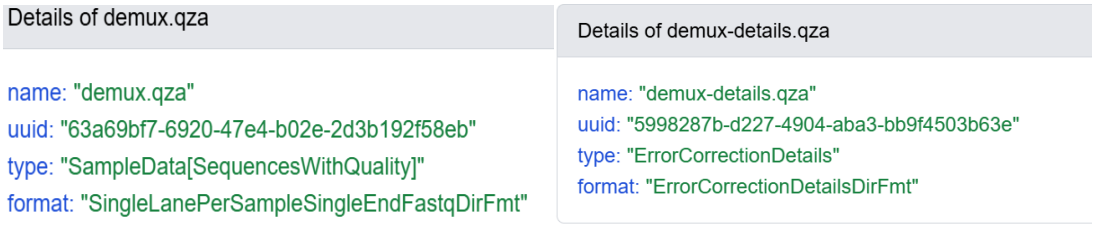


Figure 2: This figure shows the results output details for Demultiplexing sequences data. The details include the name, uuid, type, and format.

Input: qiime demux summarize, a sequence of many DNA samples were sequenced together in one run. This input helps with evaluating the data and to remove samples with few readings.

**--i-data demux.qza **

--o-visualization demux.qzv

Output: demux.qzv file that has counts summary for 34 forward sequences and forward reads frequency histogram

Results:

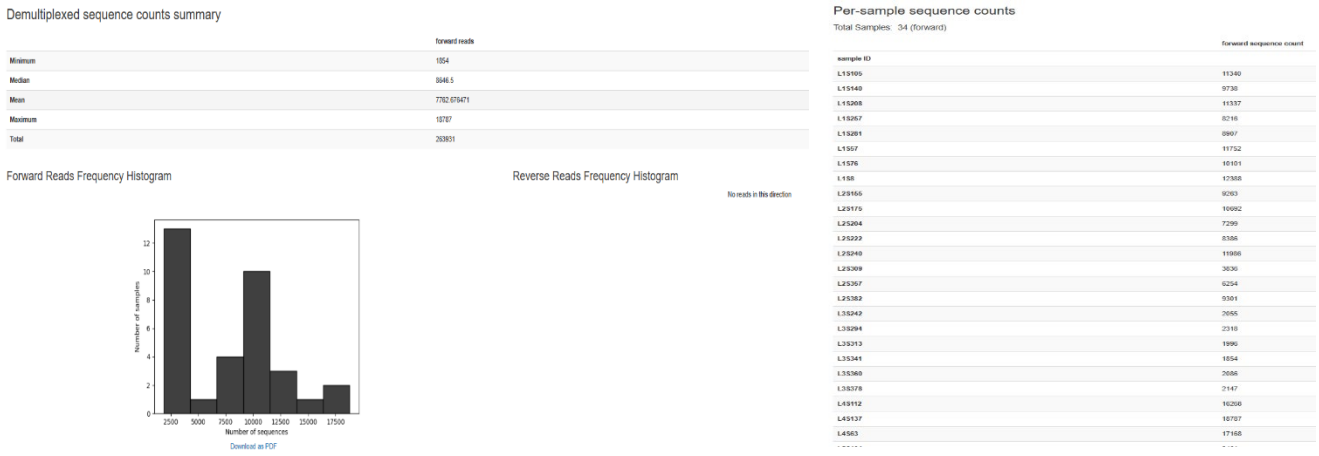


Figure 3: shows the results for the forward reads. The minimum value indicates the lowest number of reads that are found in the single sample. Median is 8646.5 reads, the value for all the samples that were count. Mean is the average number of reads per sample = 7763. Maximum is the highest number of reads in the sample=18787. The total number of reads across all samples= 263931 reads. The Histogram presents the frequency of the most samples cluster around mean/median. The histogram helps with knowing the lowest and highest sample to help with analysis.

Sequence quality control and feature table construction

Option 1: DADA2

- **Input:**
  qiime dada2 denoise-single \ ➔
  --i-demultiplexed-seqs demux.qza \ ➔
  --p-trim-left 0 \ ➔ trimming the starts of the sequence at base 1.
  --p-trunc-len 120 \ ➔ truncates to read all 120 bp.

--o-representative-sequences rep-seqs.qza \ ➔
--o-table table.qza \ ➔ I think to show how many times each ASV appears
--o-denoising-stats stats.qza➔ the stat to record the stats per sample.

Output: 3 files:

- **Rep- seqs.qza**
- **Table.qza**
- **State.qza**

**qiime metadata tabulate **
**--m-input-file stats.qza **
--o-visualization stats.qzv ➔ to visualize the stats.qzv file

Figure 4: This figure shows the results output details for Sample metadata The details include the name, uuid, type, and format.

Sample metadata: this a google sheet data that consist of samples data. I used this

[wget \

-O "sample-metadata.tsv" \

"https://data.qiime2.org/2024.10/tutorials/moving-pictures/sample_metadata.tsv"]

This file “Sample metadata” is uses for the rest of the tutorial

---

Obtaining and importing data

Input: mkdir emp-single-end-sequences ➔ the purpose of that code to make a directory

Input: qiime tools import \

--type EMPSingleEndSequences \ ➔ this code for contain sequences that are multiplexed, meaning that the sequences have not yet been assigned to samples.

Details of rep-seqs.qza

name: "rep-seqs.qza"
uuid: "650bf640-f8a7-4277-a3c0-e597d6298c01"
type: "FeatureData[Sequence]"
format: "DNASequencesDirectoryFormat"

Details of stats.qza

name: "stats.qza"
uuid: "02992e91-42a1-442e-afa4-de6b6aa3b195"
type: "SampleData[DADA2Stats]"
format: "DADA2StatsDirFmt"

Details of table.qza

name: "table.qza"
uuid: "8bd3abfb-2f0f-40e2-ada6-5d732ac7d67a"
type: "FeatureTable[Frequency]"
format: "BIOMV210DirFmt"

--input-path emp-single-end-sequences \ ➔ to make a new folder for the raw data

--output-path emp-single-end-sequences.qza ➔ to import .qza file

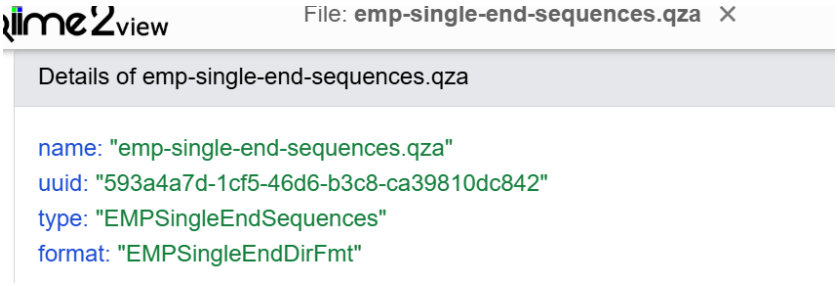


Figure 5: This figure shows the results output details for emp-single-end-sequences. The details include the name, uuid, type, and format.

-This code used to check [qiime tools peek emp-single-end-sequences .qza] the UUID, type, and format for the newly-imported sequences.

Output:

stdout:

```
UUID:        593a4a7d-1cf5-46d6-b3c8-ca39810dc842
Type:        EMPSingleEndSequences
Data format: EMPSingleEndDirFmt
```

Figure 6: it shows the UUID, Type, and Data format for the emp-single-end-sequences.qza using the qiime tools peek.

Demultiplexing sequences: to know which sequences is associated with each sample.

Input: qiime demux emp-single \ ➔ EMP single-end demultiplexing method

--i-seqs emp-single-end-sequences.qza \ ➔ sequences have sample barcodes that need to be matched with metadata.

--m-barcodes-file sample-metadata.tsv \ ➔ This is a TSV file mapping each sample ID to its barcode sequence

--m-barcodes-column barcode-sequence \ ➔ Tells QIIME which column in the metadata contains the barcodes

--o-per-sample-sequences demux.qza \ ➔ the **main output**

--o-error-correction-details demux-details.qza ➔ This file shows any barcode correction attempts that were made.

Output: 2 files named as demux-details.qza and demux.qza

Results:

Figure 7: it shows the UUID, Type, and Data format for demux-details.qza and demux.qza. The demux.qza contain contains sequences assigned to each sample after barcode-based demultiplexing. demux-details.qza shows barcode error correction and matching accuracy. These files confirm successful separation of samples for downstream analysis.

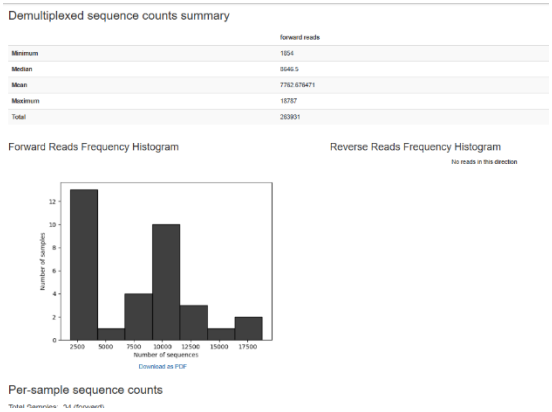
Input:

qiime demux summarize \

--i-data demux.qza \

--o-visualization demux.qzv➔ this code for visualization of the file.

Output: one file called demux.qzu -the file consist of minimum, maximum, mean, and total.



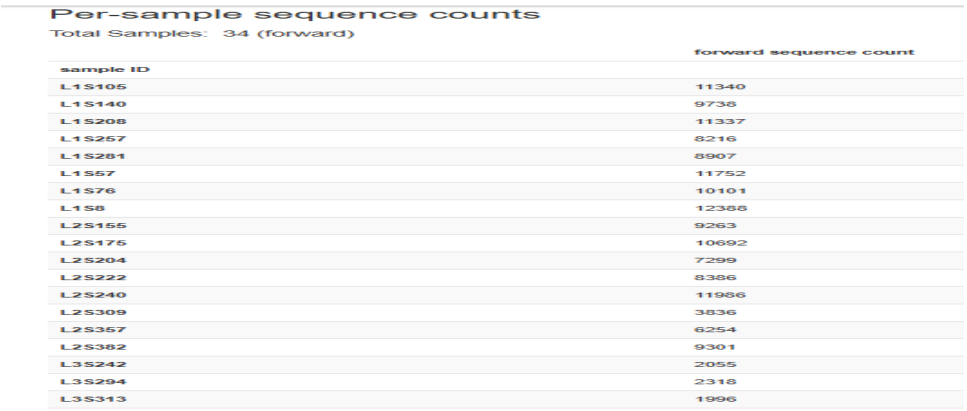


Figure 7: is showing the
results for demux.qzu. its include the results for the forward reads. The minimum value indicates the lowest number of reads that are found in the single sample. Median is 8646.5 reads, the value for all the samples that were count. Mean is the average number of reads per sample = 7763. Maximum is the highest number of reads in the sample=18787. The total number of reads across all samples= 263931 reads. The Histogram presents the frequency of the most samples cluster around mean/median. The histogram helps with knowing the lowest and highest sample to help with analysis.

Sequence quality control and feature table construction¶

Option 1: DADA2 → it's a pipeline for detecting and correcting illumina amplicon sequence.

Input: qiime dada2 denoise-single \ ➔ to denoising single

--i-demultiplexed-seqs demux.qza \ ➔ sample-separated sequence

--p-trim-left 0 \ ➔ Remove bases from start of reads ➔ remove primers or low-quality bases from the beginning

--p-trunc-len 120 \ ➔ cut reads at 120pb to remove poor-quality tails.

--o-representative-sequences rep-seqs-dada2.qza \ ➔ **representative sequences (ASVs)**, which are cleaned, denoised, and unique sequences.

--o-table table-dada2.qza \ ➔ to show table

--o-denoising-stats stats-dada2.qza ➔ to show statistics per sample.

Output: 3 files

- stats-dada2.qz, table-dada2.qza, and rep-seqs-dada2.qza.

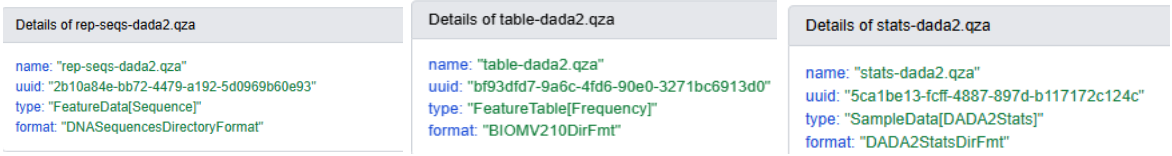


Figure 8: it shows the UUID, Type, and Data format for stats-dada2.qz, table-dada2.qza, and rep-seqs-dada2.qza.

Option 2: Deblur

Input: qiime quality-filter q-score \

--i-demux demux.qza \ ➔ demultiplexed reads

--o-filtered-sequences demux-filtered.qza \ ➔ sequences that passed quality filtering

--o-filter-stats demux-filter-stats.qza ➔ statistics showing

Output:2 files demux-filtered.qza and demux-filter-stats.qza

Figure 9: it shows the UUID, Type, and Data format for demux-filtered.qza and demux-filter-stats.qza

---

qiime deblur denoise-16S method ➔ to apply Deblur workflow. That requires one parameter. This method **denoises** sequences to identify high-quality, biologically meaningful features known as **Amplicon Sequence Variants (ASVs**

**qiime deblur denoise-16S **

**--i-demultiplexed-seqs demux-filtered.qza **

--p-trim-length 120 ** ➔ trims all sequences to exactly 120 base pairs.**

**--o-representative-sequences rep-seqs-deblur.qza ** ➔ taxonomy assignment and downstream analysis

--o-table table-deblur.qza ** ➔ how many times each ASV occurs in each sample.**

**--p-sample-stats ** ➔ how many sequences were input.

--o-stats deblur-stats.qza ➔ summary file with those per-sample stats

Output: 3 files deblur-stats.qza, table-deblur.qza, and rep-seqs-deblur.qza

rep-seqs-deblur.qza to representative ASVs (clean, unique sequences)

table-deblur.qza to show feature table showing ASV counts per sample

deblur-stats.qza per-sample stats (input reads, retained, discarded, etc.)

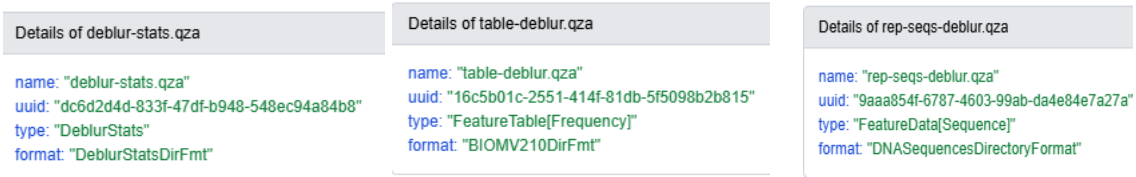


Figure 10: it shows the UUID, Type, and Data format for **deblur-stats.qza, table-deblur.qza, and rep-seqs-deblur.qza.**

Input: qiime metadata tabulate \ ➔ to create visual table

--m-input-file stats-dada2.qza \ ➔ containing summary statistics

--o-visualization stats-dada2.qzv ➔ output visualization fil

Output: stats-dada2.qzv

This file won't necessarily reflect dynamic sorting or filtering options based on the interactive table below.

Search:

sample-id #q2:types	input numeric	filtered numeric	percentage of input passed filter numeric	denoised numeric	non-chimeric numeric	percentage of input non-chimeric numeric
L1S105	11340	8571	75.58	8499	7780	68.61
L1S140	9738	7677	78.84	7605	7163	73.56
L1S208	11337	9261	81.69	9152	8152	71.91
L1S257	8216	6705	81.61	6627	6388	77.75
L1S281	8907	7067	79.34	6976	6615	74.27
L1S57	11752	9299	79.13	9260	8702	74.05
L1S76	10101	8395	83.11	8337	7867	77.88
L1S8	12388	7663	61.86	7624	7033	56.77
L2S155	9263	4112	44.39	3932	3932	42.45
L2S175	10692	4546	42.52	4386	4386	41.02
L2S204	7299	3379	46.29	3199	3158	43.27
L2S222	8386	3485	41.56	3187	3187	38
L2S240	11986	5183	43.24	5094	5061	42.22
L2S309	3836	1550	40.41	1419	1419	36.99
L2S357	6254	2526	40.39	2373	2373	37.94
L2S382	9301	4279	46.01	4202	4089	43.96
L3S242	2055	970	47.2	951	897	43.65
L3S294	2318	1313	56.64	1225	1225	52.85
L3S313	1996	1191	59.67	1103	1103	55.26
L3S341	1854	1109	59.82	962	962	51.89
L3S360	2086	1132	54.27	969	969	46.45
L3S378	2147	1358	63.25	1322	1249	58.17
L4S112	16268	8603	52.88	8362	8340	51.27
L4S137	18787	10064	53.57	9921	9820	52.27
L4S63	17168	10096	58.81	9744	9744	56.76
L5S104	3461	2253	65.1	2227	2227	64.35
L5S155	2493	1828	73.33	1800	1800	72.2

Figure 11: his figure shows the summary of denoising statistics for each sample after processing with the DADA2 plugin in QIIME 2. Each row corresponds to a sample, with columns showing the number of input reads, reads retained after filtering, denoised reads, and final non-chimeric reads. The percentages help assess how much of the input data was retained through each step. Samples with higher retention, like L1S208 and L1S140, indicate better quality data, while those with lower percentages, such as L2S175, may require closer inspection or filtering. Overall, this table provides a snapshot of the effectiveness of the quality control and denoising pipeline for each sample in the dataset.

FeatureTable and FeatureData summaries :" After the quality filtering step completes, you'll want to explore the resulting data"

Input:

qiime feature-table summarize \ ➔ action to summarize a feature table

--i-table table.qza \ ➔ containing feature counts

--o-visualization table.qzv \ ➔ visualization file

--m-sample-metadata-file sample-metadata.tsv ➔ sample metadata file to link sample IDs with details

qiime feature-table tabulate-seqs \ ➔ generate a table of all sequences

--i-data rep-seqs.qza \ ➔ DNA sequences representing each unique feature (ASV).

--o-visualization rep-seqs.qzv ➔ table showing each ASV's ID and its sequence.

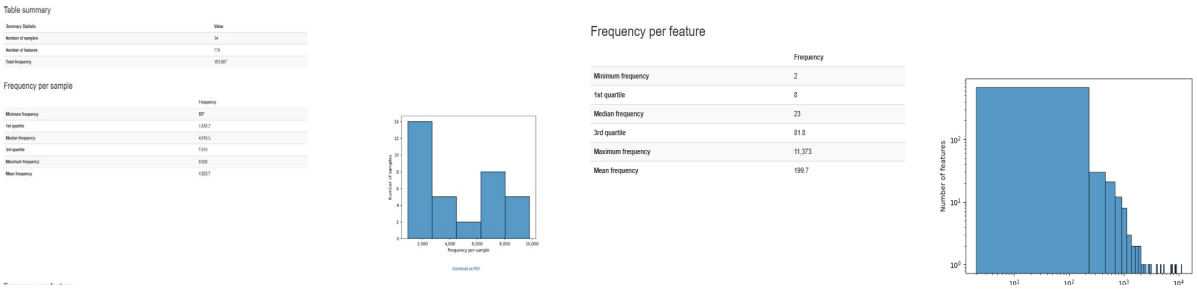


Figure 11: This figure presents key insights from the feature table statistics and frequency histograms generated in QIIME 2. The dataset contains 34 samples and 770 unique features (ASVs), with a total of 153,807 reads, indicating a moderately rich and diverse microbial community. Sample sequencing depth varies, with some samples having significantly higher or lower read counts, which impacts rarefaction decisions in diversity analyses. Feature frequency distribution reveals that while a few ASVs are highly abundant, most are present at low frequencies, reflecting the common pattern of microbial communities dominated by a few taxa. The histograms visually support these findings, highlighting both inter-sample variation and the skewed distribution of feature abundance.

Generate a tree for phylogenetic diversity analyses "QIIME supports several phylogenetic diversity metrics, including Faith's Phylogenetic Diversity and weighted and unweighted UniFrac"

Input: qiime phylogeny align-to-tree-mafft-fasttree \ ➔ sequence alignment, masking, and phylogenetic tree construction

--i-sequences rep-seqs.qza \ ➔ Specifies the input artifact containing the representative sequences

--o-alignment aligned-rep-seqs.qza \ ➔ This step uses the MAFFT algorithm to align

--o-masked-alignment masked-aligned-rep-seqs.qza \ ➔ contains the masked alignment

--o-tree unrooted-tree.qza \

--o-rooted-tree rooted-tree.qza ➔ contains the rooted phylogenetic tree

- Output: 4 files : aligned-rep-seqs.qza, masked-aligned-rep-seqs.qza, rooted-tree.qza, and unrooted-tree.qza

Details of unrooted-tree.qza

name: "unrooted-tree.qza"
uuid: "5ce076fa-2607-4bcf-9e03-1be48b8b4346"
type: "Phylogeny[Unrooted]"
format: "NewickDirectoryFormat"

Details of rooted-tree.qza

name: "rooted-tree.qza"
uuid: "8e4d0c0b-893b-4bb3-86a5-6c85228f20a7"
type: "Phylogeny[Rooted]"
format: "NewickDirectoryFormat"

Details of masked-aligned-rep-seqs.qza

name: "masked-aligned-rep-seqs.qza"
uuid: "5b1cdbda-fb60-4a66-922f-73e81a3975c3"
type: "FeatureData[AlignedSequence]"
format: "AlignedDNASequencesDirectoryFormat"

Details of aligned-rep-seqs.qza

name: "aligned-rep-seqs.qza"
uuid: "2598843a-9454-4d39-8f77-d68fc14396e1"
type: "FeatureData[AlignedSequence]"
format: "AlignedDNASequencesDirectoryFormat"

Figure 12: it shows the UUID, Type, and Data format for aligned-rep-seqs.qza, masked-aligned-rep-seqs.qza, rooted-tree.qza, and unrooted-tree.qza.

Alpha and beta diversity analysis

qiime diversity core-metrics-phylogenetic \ ➔ for diversity.

--i-phylogeny rooted-tree.qza \ ➔ containing the rooted phylogenetic tree

--i-table table.qza \ ➔ feature table

--p-sampling-depth 1103 \ ➔ sampling depth

--m-metadata-file sample-metadata.tsv \ ➔ map sample-specific attributes

--output-dir core-metrics-results ➔ to specifies the output directory.

Output: the output are multiple files:

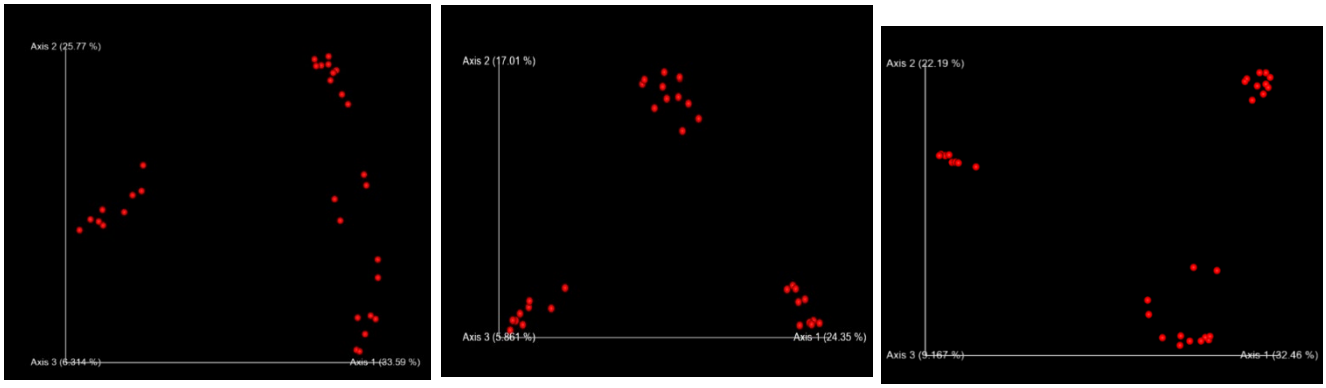


Figure 13: Beta Diversity Visualization Using Four Distance Metrics. It represent Principal Coordinates Analysis (PCoA) plots, each illustrating differences in microbial community composition across samples using a distinct beta diversity metric. Unweighted UniFrac and Jaccard plots highlight differences based solely on presence or absence of taxa, with the former incorporating phylogenetic relationships and the latter not. Weighted UniFrac and Bray–Curtis plots incorporate abundance data, with Weighted UniFrac also considering phylogenetic relatedness. Samples that cluster together in the plots share more similar microbial communities, while greater separation indicates higher dissimilarity. These visualizations help interpret how microbial communities vary between conditions, such as body sites or treatments, by revealing patterns in both community membership and structure.

Alpha rarefaction plotting: alpha diversity as a function of sampling depth.

Input: qiime diversity alpha-rarefaction \ ➔ generates a plot to explore

--i-table table.qza \ ➔ counts of observed features

--i-phylogeny rooted-tree.qza \ ➔ compute certain diversity metrics --p-max-depth 4000 \ ➔ subsample up to 4000 sequences

--m-metadata-file sample-metadata.tsv \ ➔ links each sample ID to its experimental metadata

--o-visualization alpha-rarefaction.qzv ➔ interactive visualization showing alpha diversity.

Output: alpha-rarefaction.qzv

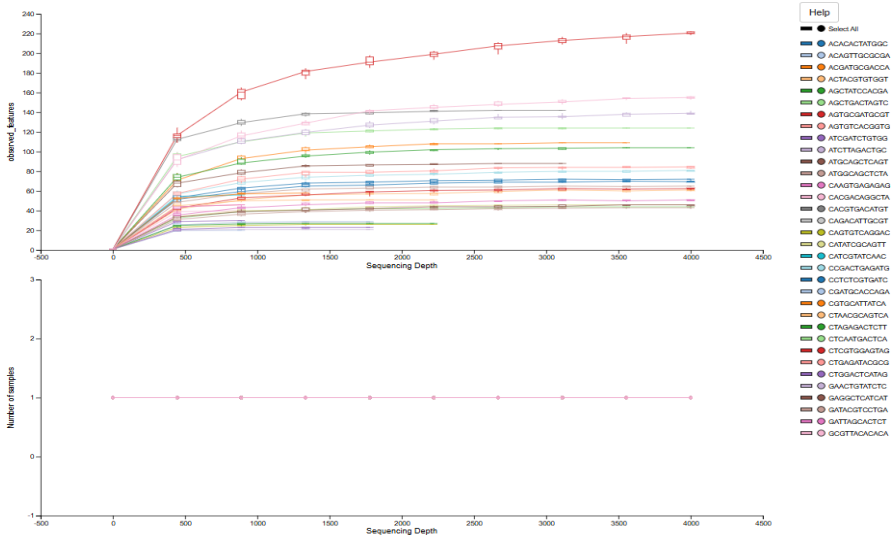


Figure 14: This figure shows alpha rarefaction curves, which illustrate how within-sample microbial diversity increases with sequencing depth, up to 4,000 sequences per sample. Samples that plateau indicate sufficient sequencing to capture most of the diversity present. The plot also compares diversity across metadata groups, helping identify biological patterns or technical biases.

Beta analysis: "sample composition in the context of categorical metadata using PERMANOVA**"**

Input: qiime diversity beta-group-significance \ ➔ perform beta diversity group significance analysis

--i-distance-matrix core-metrics results/unweighted_unifrac_distance_matrix.qza \ ➔specifies the input distance matrix

--m-metadata-file sample-metadata.tsv \ ➔ option specifies the sample metadata file

--m-metadata-column body-site \ ➔ These options specify which metadata column

--o-visualization core-metrics-results/unweighted-unifrac-body-site-significance.qzv \ ➔ flags specify the output file

--p-pairwise ➔ pairwise comparisons between groups will be performed.

- **Output: 2 files** core-metrics-results/unweighted-unifrac-body-site-significance.qzv.
- **core-metrics-results/unweighted-unifrac-subject-group-significance.qzv**

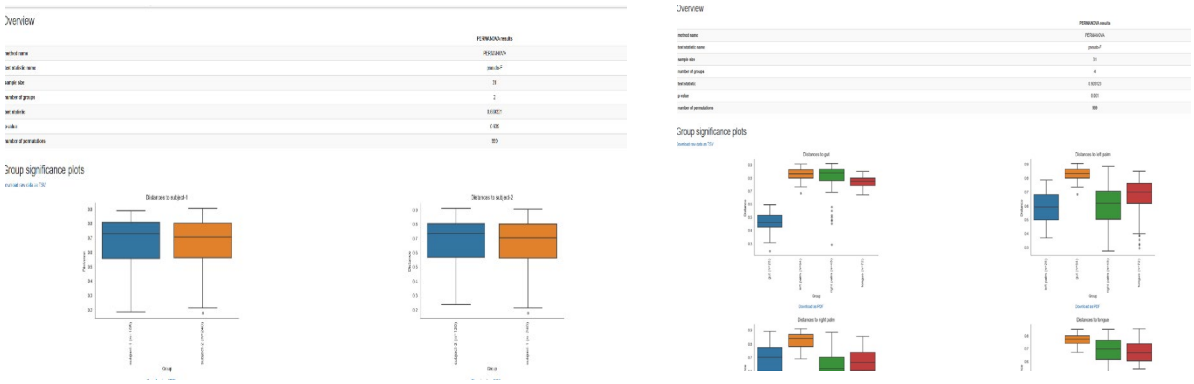


Figure 15: figure shows the results of a statistical test for beta diversity differences between body-site groups based on the Unweighted UniFrac distance matrix. The plot displays the significance of the differences in microbial community composition across body-sites, with p-values indicating whether the differences between groups are statistically significant. Beta-diversity plots show microbial community differences based **on body-site or antibiotic use**, with the first principal component explaining the most variation. Alpha-diversity metrics reveal richness differences across body sites and antibiotic use, while barplots and taxonomic analyses highlight microbial taxa patterns

---

Taxonomic analysis "to explore the taxonomic composition of the samples"

Input: qiime feature-classifier classify-sklearn \ ➔ to assign taxonomy for the sequences

--i-classifier gg-13-8-99-515-806-nb-classifier.qza \ ➔ trained on the Greengenes database

--i-reads rep-seqs.qza \ ➔ representative sequences

--o-classification taxonomy.qza ➔ taxonomy assignments

qiime metadata tabulate \ ➔ easy-to-read table visualization

--m-input-file taxonomy.qza \ ➔ output taxonomy file

--o-visualization taxonomy.qzv ➔ result as an interactive

Figure 15: This figure shows the taxonomic assignments of representative sequences using a pre-trained Naive Bayes classifier (gg-13-8-99-515-806-nb-classifier.qza) based on the Greengenes 13_8 99% OTUs database. The resulting taxonomy file taxonomy.qza contains hierarchical classifications from kingdom to genus level for each ASV. This classification provides insight into the microbial composition of the samples. The interactive visualization taxonomy.qzv allows users to explore the identities and distribution of taxa across all samples.

To view taxonomic composition

Input: qiime taxa barplot \

--i-table table.qza \

--i-taxonomy taxonomy.qza \

--m-metadata-file sample-metadata.tsv \

--o-visualization taxa-bar-plots.qzv

Output: taxa-bar-plots.qzv

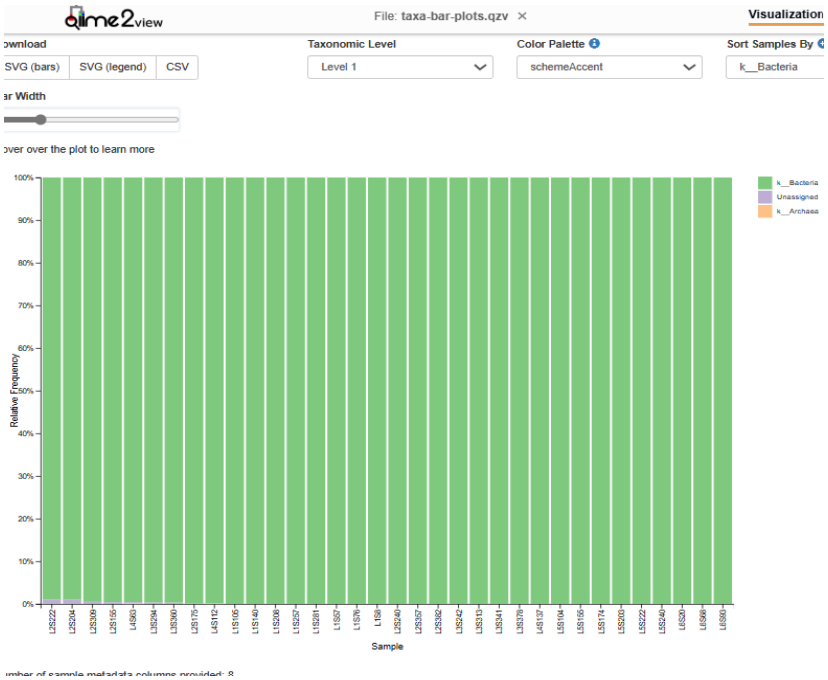


Figure 16: **Figure Legend: Taxonomic Composition Bar Plots**

This figure shows the relative abundance of microbial taxa across all samples, visualized as stacked bar plots. Each color represents a different taxonomic group, and bars are grouped by sample IDs, allowing for comparison of microbial community composition between samples. The plot is

interactive, enabling exploration at different taxonomic levels (e.g., phylum, genus) by adjusting the display settings. This visualization helps identify dominant taxa and observe patterns related to sample metadata such as body site or treatment.

Differential abundance testing with ANCOM-BC : ompositionally-aware linear regression model that allows for testing differentially abundant features across groups while also implementing bias correction.

Input: qiime feature-table filter-samples \ ➔ to filter samples from a feature table

--i-table table.qza \ ➔ input feature table

--m-metadata-file sample-metadata.tsv \ ➔ contains sample information

--p-where "[body-site]='gut'" \ ➔ parameter specifies the condition

--o-filtered-table gut-table.qza ➔ samples with the body-site 'gut'

Output: gut-table.qza

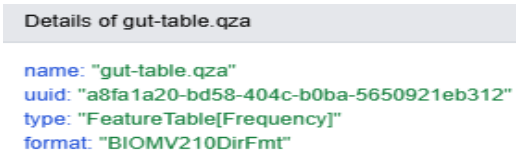


Figure 17: This figure shows the filtered feature table (gut-table.qza) that includes only samples from the "gut" body-site, as specified in the metadata file. It contains the abundance of features across the selected gut samples, enabling focused analysis of the gut microbiome.
