Amplicon Sequence Analysis Pipeline (ASAP2)

Contents

Overview

Amplicon sequencing of marker genes such as 16S rDNA, 18S rDNA, ITS and others has been widely used to survey and characterize microbial communities in countless ecological and environmental studies. However, the complex data analyses have required many interfering manual steps often leading to inconsistencies in results. Here, we have developed a pipeline, Amplicon sequence analysis pipeline 2 (ASAP 2), to automate and glide through the processes without the usual manual inspections and user's interference, for instance, in the detection of barcode orientation, selection of high-quality region of reads, and determination of resampling depth and many more. The pipeline integrates all the analytical processes such as importing data, demultiplexing, summarizing read profiles, trimming quality, denoising, removing chimeric sequences, and making the feature (ASV) table among others, using QIIME 2, Vegan and other tools. The pipeline accepts multiple file formats as input including multiplexed or demultiplexed, paired-end or single-end, barcode inside or outside and raw or intermediate data (e.g. feature table). The outputs include taxonomic classification, alpha/beta diversity, community composition, ordination analysis and statistical tests (variable selection, CCA, RDA, etc.). ASAP 2 supports merging multiple sequencing runs which helps integrate and compare data from different sources (public databases and collaborators). The pipeline minimizes hands-on interference and runs amplicon sequencing analysis automatically and consistently.

Frequently Asked Question

What analysis does it do?

It covers all the process involved in QIIME 2 including data import, demultiplexing, read profile summarization, high-quality region detection, quality trimming, denoising (using DADA2), removing chimeric sequence, making feature (ASV) table, taxonomic classifcation, alpha/beta diversity, alpha rarefaction, community composition etc.

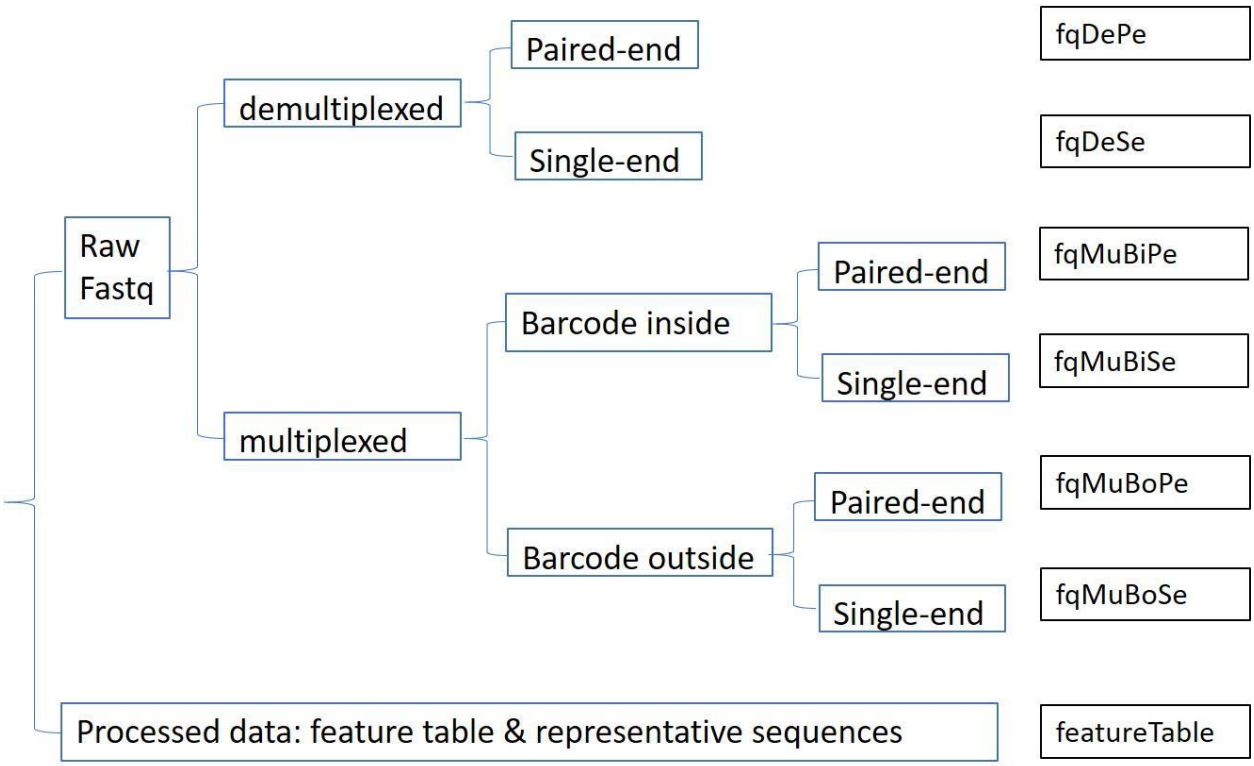
Further, it does common statistical analysis such as group significance tests on alpha diversity answering the question what factors impact the community diverstiy. It also does ordination analysis including canonical correspondence analysis (CCA), redundancy analysis (RDA) and variable selection analysis, which answer the question what environmental factors significantly affect the community composition, and which species are correlated to them.

Who can use it?

This pipeline aims to help scientists with no or limited bioinformatics skills including Linux and programming. For people who master bioinformatics skills but don't want to spend time in reading the long QIIME 2 tutorial and deciding tons of commands, this pipeline helps save a lot of time.

What data formats are accepted?

ASAP2 accepts many sequence data formats, including demultiplexed and multiplexed, barcode-inside and barcode-outside, paired-end and single-end, and even processed data with feature table as input. Below are the data formats that are accepted to ASAP2.



Preparation of Input Data

Metadata

Prepare your metadata using the template of QIIME 2:

https://docs.google.com/spreadsheets/d/1zCErEpDf5JCXoMmDQWYToIPn07vWdsyKg4UbFc2JgLA/edit#gid=0

Users can put grouping information and environmental factors information as columns for downstream analysis such as PCoA.

Specially, for multiplexed barcode inside data, add a column 'barcode-sequence2' (will be detected by the pipeline) if there are barcodes at both ends to include double checking of barcodes in the demultiplexing step.

It is important to validate your metadata format using Keemei (https://keemei.qiime2.org/, install first) because incorrect format or naming will cause interruption of the pipeline. After validation, name the file as metadata.tsv.

Fastq Read Data

Demultiplexed paired-end (fqDePe)

Demultiplexed data means that the sequences have been assigned to samples and each sample has its own file(s). Paired-end data has forward and reverse read files, usually tagged with R1 and R2, respectively. Demultiplexed data usually does not have barcode in sequence and barcode file.

Each sample has forward and reverse read files. Name them as BAQ1552_67_L001_R1_001.fastq.gz and BAQ1552_67_L001_R2_001.fastq.gz. BAQ1552 is the sample name and you should replace it with your own. The strings 67, L001, and 001 do not really matter. If you don't have them, just use the numbers of this example.

Put the paired-end files of each sample in seqData/. Put the seqData/ and the validated metadata.tsv in a folder **fqDePe-project_cas1**. Replace project_cas1 with your own project name.

Demultiplexed single-end (fqDeSe)

Demultiplexed data means that the sequences have been assigned to samples and each sample has its own file(s). Single-end data has only forward read file, usually tagged with R1. Demultiplexed data usually does not have barcode in sequence and barcode file.

Name the files as BAQ1552_67_L001_R1_001.fastq.gz. BAQ1552 is the sample name and you should replace it with your own. The strings 67, L001, and 001 do not really matter. If you don't have them, just use the numbers of this example.

Put the single-end files of each sample in seqData/. Put the seqData/ and the validated metadata.tsv in a folder **fqDeSe-project_cas1**. Replace project_cas1 with your own project name.

Multiplexed barcode-inside paired-end (fqMuBiPe)

Multiplexed data means that the sequences of multiple samples are in the same files and are discriminated with barcodes. Barcode-inside means the barcode sequences are linked to the forward or reverse or both reads in read file(s) and there is no separate barcode file(s). Paired-end data has forward and reverse read files, usually tagged with R1 and R2, respectively. The barcodes

in forward reads are mandatory (specified as barcode-sequence column in metadata file). If you have dual-indexing data and want to remove sequences with unmatched barcodes, put reverse barcodes in the metadata file in the column barcode-sequence2.

Name the paired-end files as forward.fastq.gz and reverse.fastq.gz and put them in seqData/. Put the seqData/ and the validated metadata.tsv in a folder **fqMuBiPe-project_emp1**. Replace project_emp1 with your own project name.

Multiplexed barcode-inside single-end (fqMuBiSe)

Multiplexed data means that the sequences of multiple samples are in the same file and are discriminated with barcodes. Barcode-inside means the barcode sequences are linked to the forward read and there is no separate barcode file. Single-end data has only forward read file and is usually tagged with R1.

Name the single-end file as sequences.fastq.gz put it in seqData/. Put the seqData/ and the validated metadata.tsv in a folder **fqMuBiSe-project_emp1**. Replace project_emp1 with your own project name.

Multiplexed barcode-outside paired-end (fqMuBoPe)

Multiplexed data means that the sequences of multiple samples are in the same files and are discriminated with barcodes. Barcode-outside means the barcode sequences are trimmed from the forward or reverse or both reads and there is separate barcode file(s), usually tagged with I1 (I1 and I2 for double barcoding). Here only one barcode file (either I1 or I2) is needed. Put the barcodes of samples in the metadata.tsv in the column barcode-sequence. The barcodes in the metadata file and the read file can be either the same or reverse complementary as the pipeline will detect the orientation automatically. Paired-end data has forward and reverse read files, usually tagged with R1 and R2, respectively.

Name the paired-end files and barcode file as forward.fastq.gz, reverse.fastq.gz and barcodes.fastq.gz, and put them in seqData/. Put the seqData/ and the validated metadata.tsv in a folder **fqMuBoPe-project_emp1**. Replace project_emp1 with your own project name.

Multiplexed barcode-outside single-end (fqMuBoSe)

Multiplexed data means that the sequences of multiple samples are in the same file and are discriminated with barcodes. Barcode-outside means the barcode sequences are trimmed from the forward read and there is a separate barcode file, usually tagged with I1. Put the barcodes of samples in the metadata.tsv in the column barcode-sequence. The barcodes in the metadata file and the read file can be either the same or reverse complementary as the pipeline will detect the orientation automatically. Single-end data has only forward read file and is usually tagged with R1.

Name the single-end file and barcode file as sequences.fastq.gz and barcodes.fastq.gz, and put them in seqData/. Put the seqData/ and the validated metadata.tsv in a folder **fqMuBoSe-project_emp1**. Replace project_emp1 with your own project name.

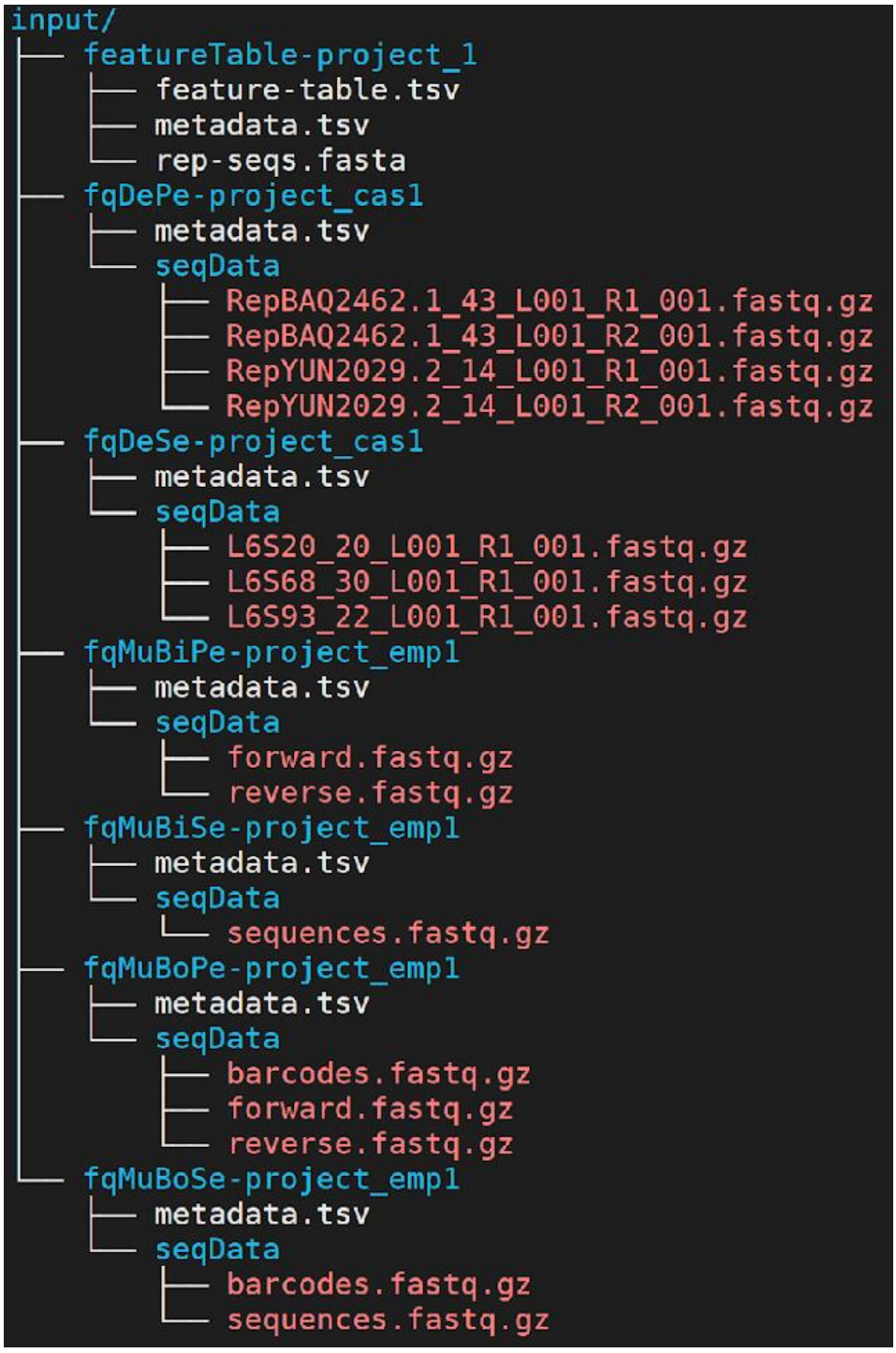
Feature Table Data

If you only have feature table and representative sequence data, name them feature-table.tsv (or feature-table.biom if it is biom format, we accept both json and hdf5 versions) and rep-seqs.fasta, respectively. If you have qza files, unzip them and get the table or sequence file in the data folder in the unzipped directory. Get the validated metadata.tsv and put the three files in a folder **featureTable-project_1**. Replace project_1 with your own project name.

Organizing Multiple Data Sets

Well done! You finally got here. Multiple projects could be gathered and analyzed in one blow.

Put all the folders in a folder (e.g. input/). It will be like this.



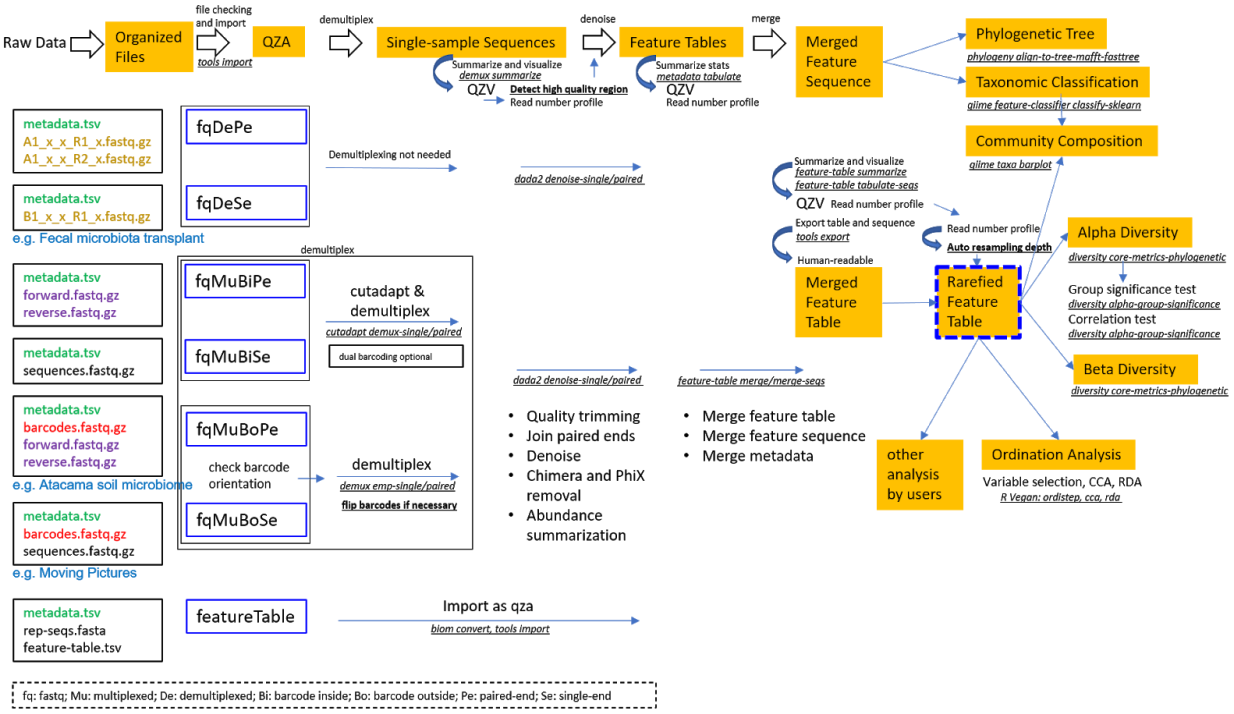
Then you can run the /bin/asap2.py to start the pipeline. Use the parameter -h for help information.

Workflow of the pipeline

The pipeline starts from sequence data in raw fastq or its compressed format (e.g. fastq and fastq.gz), or feature table with representative sequences. The sequence files can contain sequences that belong to multiple samples that were barcoded or indexed, pooled and sequenced together (multiplexed) or to only a single sample (demultiplexed). For the multiplexed data, the barcode sequence can be either inside or outside the read files. For all the above fastq data files, both single-end and paired-end formats are supported. The pipeline first converts the organized input data to QIIME Zipped Artifacts (qza) format. Multiplexed data will be demultiplexed into single-sample data. For the fastq files with barcode data inside or barcode-inside fastq files, barcodes will be cut out for demultiplexing with the available option of dual barcode validation. For the fastq files with the barcode data in separate fastq files, or barcode-outside fastq files, barcodes in the metadata file and the sequence file will be compared to determine the orientation for a correct demultiplexing.

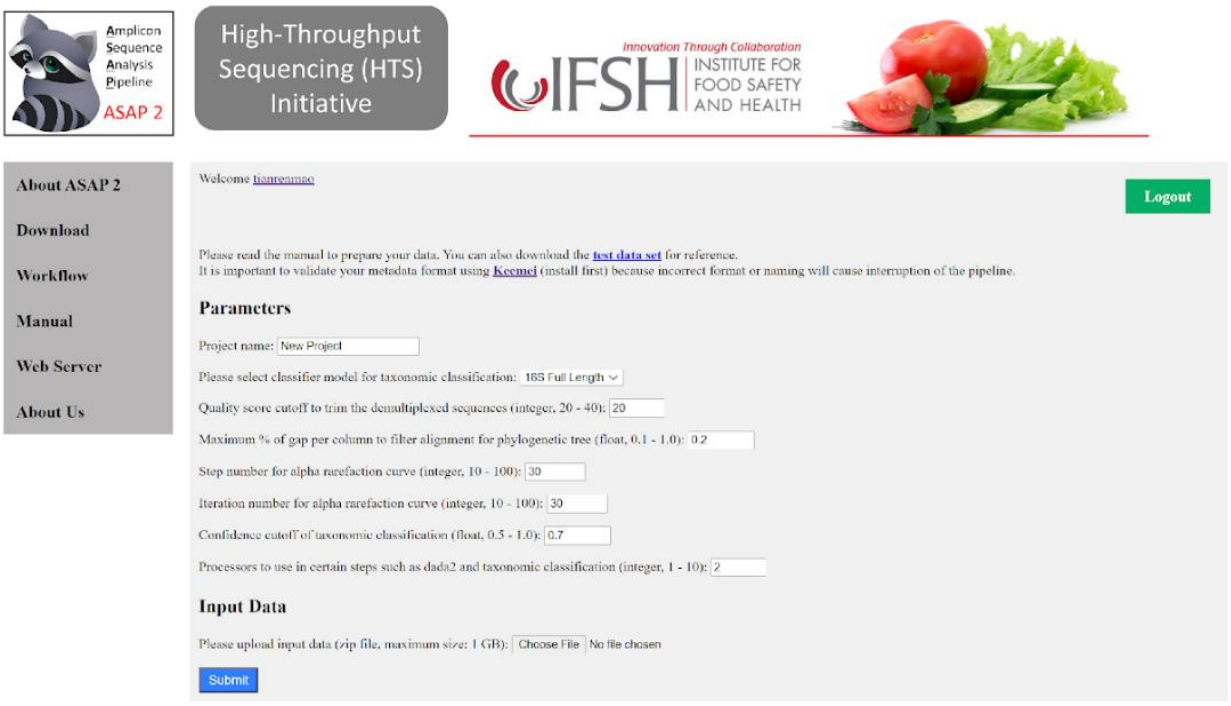
Optimal high-quality region of the reads will be identified and selected as an input parameter for the downstream denoising steps through quality trimming, paired ends joining, sequencing error correction, and removing of chimera and PhiX sequences. The high-quality reads will then be summarized to generate a feature table with information about the abundance of amplicon sequence variants (ASV) in the samples. The generated feature tables and ASV sequences of multiple projects (if applicable), together with other additional input feature tables and ASV sequences, will be merged for the purpose of combination or comparison.

Resampling will be performed on the merged feature table to eliminate the bias caused by uneven sample sizes. The merged sequences will be used for phylogenetic tree reconstruction and taxonomic classification. The output of taxonomic classification with the feature table will be used for community composition analysis. Alpha diversity and beta diversity analysis will be performed to provide information about intra- and inter-sample diversity. A group significance test and a correlation test will be performed to show the effects of environmental factors or treatments on the alpha diversity of samples. Ordination analysis will be performed using the R package Vegan to provide insights about interactions between environmental factors and species in the community.



Web server interface

A web server is also prepared to host ASAP 2 in a high-performance computer (HPC) with 20 cores CPUs (Intel Xeon CPU E5-2660 v3 @ 2.60 GHz, 40 threads) and 512 GB memory. The website front end is written in HTML, CSS and JavaScript, and the back end is written in Django. The ASAP 2 website users need to organize their data based on the data formats and submit them with their desired parameters. A typical dataset (100 samples, 1 Gbp) takes about 20 minutes using 10 threads in the web server.



Demonstration results

Using the demonstration data set Atacama soil microbiome from QIIME 2 tutorial (https://docs.qiime2.org/2020.8/tutorials/atacama-soils/), ASAP 2 completed the analysis of the organized data and parameter input under no human supervision. It determined the orientation of the barcode as reverse complement in the metadata file and the barcode sequence file. It selected the optimal high-quality region after demultiplexing, and determined the resampling depth after generating the feature table.

After demultiplexing, a table for read number of samples will be generated and saved in the file *readSummary/read_summary_demux.tsv*.

		forward sequence count	reverse sequence count
fqMuBoPe-project_emp1	YUN3428.2	1442	1442
fqMuBoPe-project_emp1	BAQ4166.1.2	1369	1369
fqMuBoPe-project_emp1	BAQ4166.2	1341	1341
fqMuBoPe-project_emp1	YUN3533.3	1321	1321
fqMuBoPe-project_emp1	BAQ3473.3	1320	1320
fqMuBoPe-project_emp1	YUN3856.3	1297	1297
fqMuBoPe-project_emp1	BAQ4166.3	1290	1290
fqMuBoPe-project_emp1	YUN3533.2	1228	1228
fqMuBoPe-project_emp1	BAQ3473.1	1148	1148
fqMuBoPe-project_emp1	YUN3259.3	1146	1146

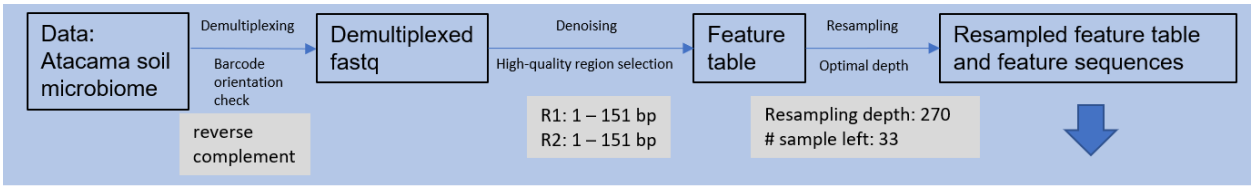
A summary showing the maximum, minimum, median and mean of read numbers will be saved in the log file.

```
fqMuBoPe-project_emp1:
# sample: 40
Max read: 1442
Min read: 439
Median read: 935
Mean read: 946
```

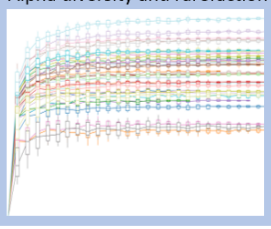
In terms of diversity, the outputs included the alpha diversity indices (number of features, Shannon index, Faith's Phylogenetic Diversity, etc.) and corresponding rarefaction. The rarefaction curves can be used to determine the sequencing saturation that indicates if the sequencing effort is sufficient or not. The effects of environmental factors on the diversity was investigated by correlation analysis. As a result, the depth and elevation were found to have a significant impact on the species diversity. Sample grouping also had an impact on the species diversity and the two groups (different transects) were identified to have a significantly different diversity. Another output was the beta diversity analysis that examines the inter-sample similarities. The result showed that the microbial communities without vegetation were clustered away from those with vegetation.

In terms of taxonomy, the representative sequences were classified at all levels. The classification results and the feature table were used to profile the community composition. A phylogenetic tree was constructed using the feature sequences, which was also used for the alpha and beta diversity analysis based on phylogenetic distance metrics.

With the community composition profile and the metadata, a variable selection analysis was performed on the same dataset to identify the environmental factors or grouping that have significant impacts on the community composition. Furthermore, a CCA and RDA analysis were used to provide insights into the correlation of environmental factors and taxa. The results showed that vegetation, depth, site and pH are the significant factors affecting the community composition and the species correlating with depth and pH can be observed from the CCA result.



Alpha diversity and rarefaction



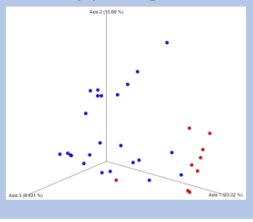
Correlation of alpha diversity and environmental factors

Shannon index – depth:
Spearman 0.37
P value 0.03
Shannon index – elevation:
Spearman 0.40
P value 0.02

Correlation of alpha diversity and groups

Group 1	Group 2	H	p-value	q-value
Baquedano (n=13)	Yungay (n=20)	4.7	0.03	0.03

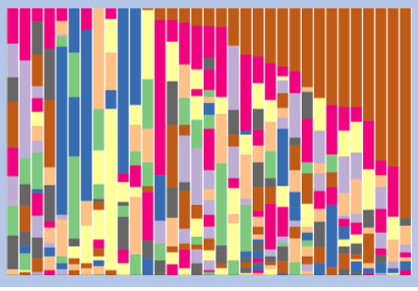
Beta diversity (unweighted UniFrac)



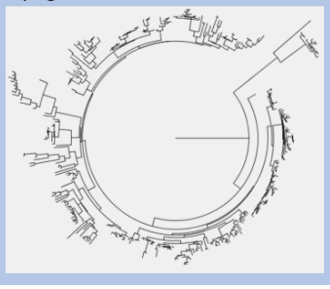
Taxonomic classification

b37b0e21bf8b789cc0122fea231e9d93	d__Bacteria; p__Actinobacteriota; c__Rubrobacteria; o__Rubrobacterales; f__Rubrobacteriaceae; g__Rubrobacter	0.99

Community composition



Phylogenetic tree



Variable selection

	Df	AIC	F	Pr(>F)
vegetation	1	263.76	2.22	0.01
depth	1	263.93	1.71	0.01
site.name	13	267.53	1.27	0.01
ph	1	266.21	1.70	0.05

CCA