Microbiome

1. Kraken2 report

Data: `/srv/common/microbiome/COLD__r09_F015__85.txt`

1. Какой это формат выходного файла?
2. Сколько микроорганизмов было определено с точностью до вида (и глубже)?
3. На какой определённый вид приходится больше всего фрагментов?
4. На какой вид батерии приходится больше всего фрагментов?
5. Какова доля фрагментов, не отнесённых с какому-либо таксону?
6. Сколько фрагментов было отнесено к роду Acetobacter, но ни к одному из конкретных видов Acetobacter?

2. Qiime2

We are using human microbiome samples from two individuals at four body sites at five timepoints, one of the individuals used antibiotics. This data came from the hypervariable region 4 (V4) of 16S rRNA and sequenced on an Illumina HiSeq.

Data: `/srv/common/microbiome/emp-single-end-sequences/sequences.fastq.gz`
`/srv/common/microbiome/emp-single-end-sequences/barcodes.fastq.gz`
Metadata: `/srv/common/microbiome/sample-metadata.tsv`

Activate environment with Qiime2

1. configure conda

```
/opt/tljh/user/bin/codna init
```

2. restart session

```
source ~/.bashrc
```

3. activate session

```
conda activate /opt/tljh/user/envs/qiime2-amplicon-2024.10
```

Input data

1. How many reads in `/srv/common/microbiome/emp-single-end-sequences/sequences.fastq.gz` file?
2. Is it single-end or paired-end fastq file?

3. Qiime uses Quiime artifacts as inputs to other commands which contain information about the type of data and the source of the data. So, the first thing we need to do is import these sequence data files into a QIIME 2 artifact

```
qiime tools import \
  --type 'EMPSingleEndSequences' \
  --input-path emp-single-end-sequences \
  --output-path emp-single-end-sequences.qza
```

Demultiplexing sequences

4. To demultiplex sequences we need to know which barcode sequence is associated with each sample. This information is contained in the sample metadata file. Find the column name in sample metadata and put it insted of `?`

```
qiime demux emp-single \
  --i-seqs emp-single-end-sequences.qza \
  --m-barcodes-file sample-metadata.tsv \
  --m-barcodes-column ? \
  --o-per-sample-sequences demux.qza \
  --o-error-correction-details demux-details.qza
```

The demux.qza QIIME 2 artifact will contain the demultiplexed sequences. The second output (demux-details.qza) presents Golay error correction details, and will not be explored in this task.

5. After demultiplexing, it's useful to generate a summary of the demultiplexing results.

```
qiime demux summarize \
  --i-data demux.qza \
  --o-visualization demux.qzv
```

We cannot easily open qiime artifcat. Let's visualise `demux.qzv` using https://view.qiime2.org/

! Provide a screenshot of forward reads frequency histogram. What number of sequences occur **second most** among samples?

QC and feature table construction (relative abundance)

6. We are using tool **DADA2** for this step. It will perform all the quality control for us and build table of relative abundance, which contains counts (frequencies) of each unique sequence in each sample in the dataset

For QC we will trim reads at the positions where quality of bases drops off (the flag `--p-trunc-len n` which truncates each sequence at position `n`)

Before running command, review the Interactive Quality Plot tab of the `demux.qzv` file in https://view.qiime2.org/. At what position the quality seems to drop off? Write the number instead of question mark

This next command may take up to 10 minutes to run, and is the slowest step in this task.

```
qiime dada2 denoise-single \
  --i-demultiplexed-seqs demux.qza \
  --p-trim-left 0 \
  --p-trunc-len ? \
  --o-representative-sequences rep-seqs-dada2.qza \
  --o-table table.qza \
  --o-denoising-stats stats.qza
```

7. Command for generating file for visualization.

```
qiime metadata tabulate \
  --m-input-file stats.qza \
  --o-visualization stats.qzv
```

What is mean number of input reads?

What is the mean number of filtered reads?

What is the mean percentage of input non-chimeric reads?

Summaries of FeatureTable and FeatureData

8. After the quality filtering step completes, the `feature-table summarize` command will give you information on how many sequences are associated with each sample and with each feature, histograms of those distributions, and some related summary statistics. The `feature-table tabulate-seqs` command will provide a mapping of feature IDs to sequences, and provide links to easily BLAST each sequence against the NCBI nt database.

```
qiime feature-table summarize \
  --i-table table.qza \
  --o-visualization table.qzv \
  --m-sample-metadata-file sample-metadata.tsv
qiime feature-table tabulate-seqs \
  --i-data rep-seqs.qza \
  --o-visualization rep-seqs.qzv
```

Visualise `table.qzv` as usual.
What is the minimum frequency per sample?

Generate tree for phylogenetic diversity analysis

9.

```
qiime phylogeny align-to-tree-mafft-fasttree \
  --i-sequences rep-seqs.qza \
  --o-alignment aligned-rep-seqs.qza \
  --o-masked-alignment masked-aligned-rep-seqs.qza \
  --o-tree unrooted-tree.qza \
  --o-rooted-tree rooted-tree.qza
```

Alpha and beta diversity analysis

10.

QIIME 2's diversity analyses are available through the `q2-diversity` plugin, which supports computing alpha and beta diversity metrics, applying related statistical tests, and generating interactive visualizations. We'll first apply the `core-metrics-phylogenetic` method, which rarefies a `FeatureTable[Frequency]` to a user-specified depth, computes several alpha and beta diversity metrics, and generates principle coordinates analysis (PCoA) plots using Emperor for each of the beta diversity metrics. The metrics computed by default are:

- Alpha diversity
  - Shannon's diversity index (a quantitative measure of community richness)
  - Observed Features (a qualitative measure of community richness)
  - Faith's Phylogenetic Diversity (a qualitative measure of community richness that incorporates phylogenetic relationships between the features)
  - Evenness (or Pielou's Evenness; a measure of community evenness)
- Beta diversity
  - Jaccard distance (a qualitative measure of community dissimilarity)
  - Bray-Curtis distance (a quantitative measure of community dissimilarity)
  - unweighted UniFrac distance (a qualitative measure of community dissimilarity that incorporates phylogenetic relationships between the features)
  - weighted UniFrac distance (a quantitative measure of community dissimilarity that incorporates phylogenetic relationships between the features)

```
qiime diversity core-metrics-phylogenetic \
  --i-phylogeny rooted-tree.qza \
  --i-table table.qza \
  --p-sampling-depth 1103 \
  --m-metadata-file sample-metadata.tsv \
  --output-dir core-metrics-results
```

Visualise `jaccard_emperor.qzv` in `core-metrics-results`. Provide a screenshot. What do you see?

3. Alpha and beta diversity

You will need to use next libraries:

```
library(microbiome)

library(knitr)

library(tidyverse)

library(MicrobeDS)

library(vegan)

library(ggplot2)

library(reshape)
```

To get the data for the tasks, use following commands in Rserver

```
alpha diversity:

            data(dietswap)

beta (intra-individual and inter-individual):

             data(peerj32)

beta (intra-overtime):

             data(MovingPictures)
```

1. calculate the alpha diversity, provide a Shannon plot and perform a Kolmogorov-Smirnov test... provide the p-value
2. provide a plot between cases and controls, showing divergance within each group samples
3. provide a timeseries of divergance over time
4. provide a plot for inter-individual divergance for each groups as a comparison