PacificBiosciences / pb-16S-nf

Stéphane Plaisance [VIB - Nucleomics Core, nucleomics@vib.be]

$ThuDec22, 2022$ - version 1.0

Contents

Introduction

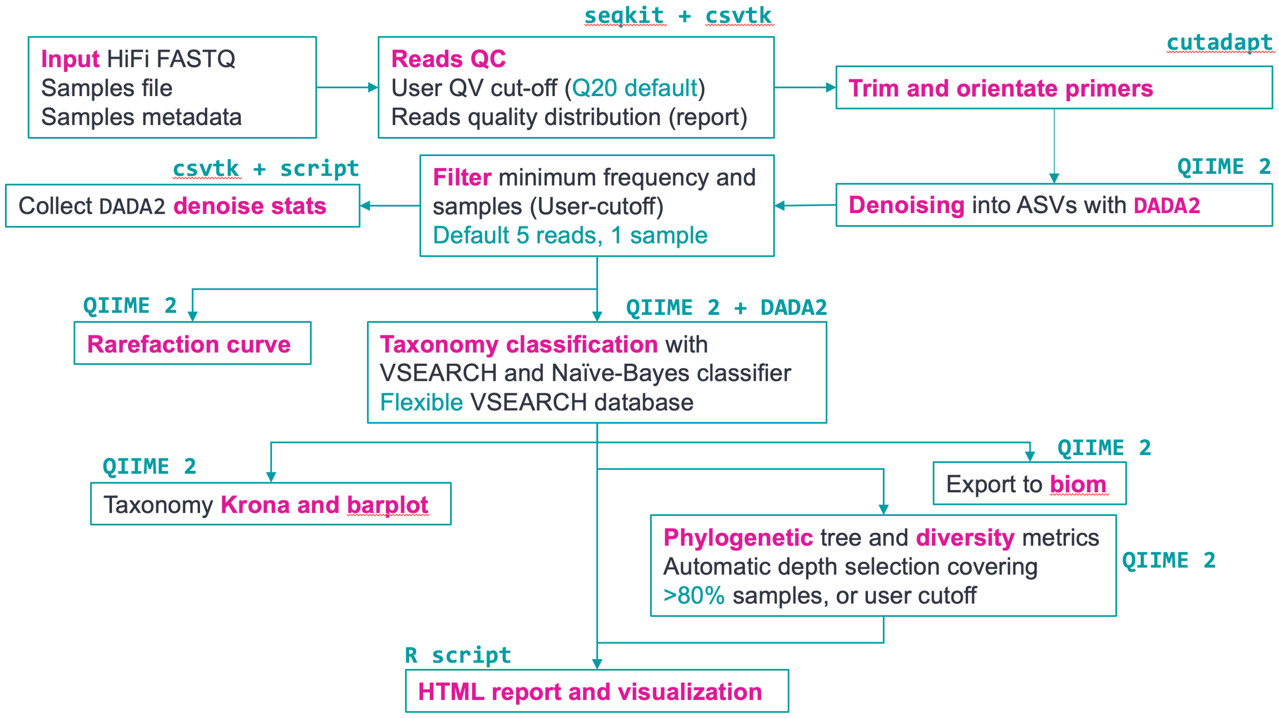
We describe here a new nextflow pipeline **pb-16S-nf** [1] developed by **Khi Pin, Chua** (@proteinosome) as part of the Pacbio open code hosted on github and can be used to analyze data obtained with the Pacbio 16S method [2]. Khi Pin is actively developing this package further and was very helpful in deploying this code and correcting a few issues.

«This Nextflow pipeline is designed to process PacBio HiFi full-length 16S data into high quality amplicon sequence variants (ASVs) using **QIIME2** and **DADA2**. It provides a set of visualization through the QIIME 2 framework for interactive plotting. The pipeline generates a HTML report for the important statistics and top taxonomies» (**taken from the github page**).

The Nextflow pipeline depends on two text files and a matching folder of demultiplexed HiFi fastq files produced by the SMRTLink platform.

The pipeline performs a number pre-processing steps followed by DADA2 and Qiime2 commands. All of it integrated and standardized for ease of use.

The general workflow is shown in the next figure

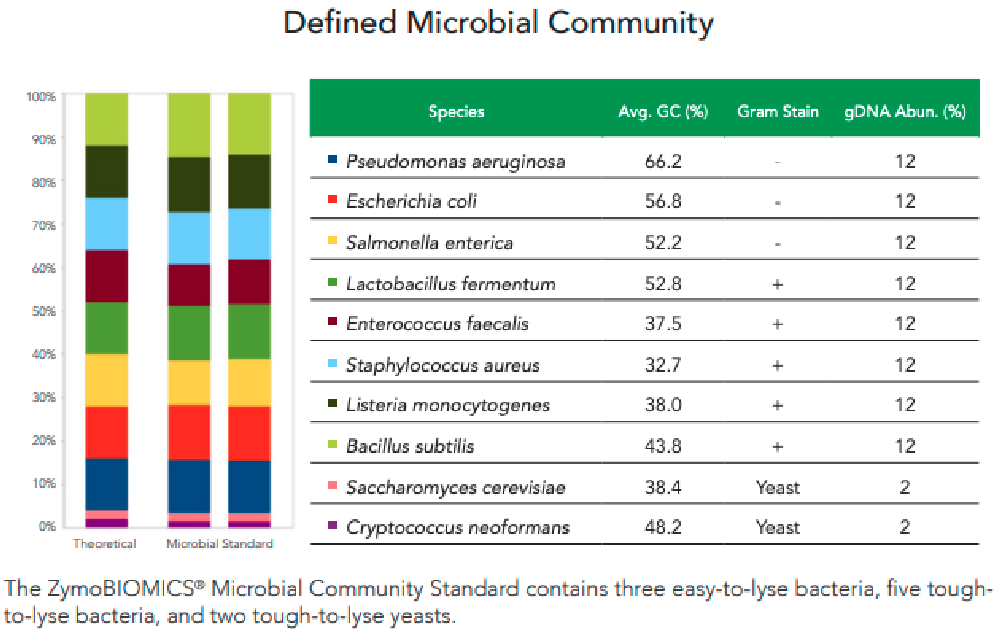

seqkit + csvtk
cutadapt
Input HiFi FASTQ
Samples file
Samples metadata
Reads QC
User QV cut-off (Q20 default)
Reads quality distribution (report)
Trim and orientate primers
csvtk + script
QIIME 2
Collect DADA2 denoise stats
Filter minimum frequency and
samples (User-cutoff)
Default 5 reads, 1 sample
Denoising into ASVs with DADA2
QIIME 2
QIIME 2 + DADA2
Rarefaction curve
Taxonomy classification with
VSEARCH and Naïve-Bayes classifier
Flexible VSEARCH database
QIIME 2
QIIME 2
Export to biom
Taxonomy Krona and barplot
Phylogenetic tree and diversity metrics
Automatic depth selection covering
>80% samples, or user cutoff
QIIME 2
R script
HTML report and visualization

[1] https://github.com/PacificBiosciences/pb-16S-nf

[2] https://www.pacb.com/wp-content/uploads/Procedure-checklist-Amplification-of-bacterial-full-length-16S-rRNA-gene-with-barcoded-primers.pdf

Theoretical composition of the Zymo mock community

In order to validate the workflow, we collected barcoded positive control samples from 5 Nucleomics Core 16S Sequel-IIe experiments and used them to compare pb-16S-nf results to the theoretical distribution present in the **Zymo mock community** [3]



Defined Microbial Community

Species	Avg. GC (%)	Gram Stain	gDNA Abun. (%)
■ *Pseudomonas aeruginosa*	66.2	-	12
■ *Escherichia coli*	56.8	-	12
■ *Salmonella enterica*	52.2	-	12
■ *Lactobacillus fermentum*	52.8	+	12
■ *Enterococcus faecalis*	37.5	+	12
■ *Staphylococcus aureus*	32.7	+	12
■ *Listeria monocytogenes*	38.0	+	12
■ *Bacillus subtilis*	43.8	+	12
■ *Saccharomyces cerevisiae*	38.4	Yeast	2
■ *Cryptococcus neoformans*	48.2	Yeast	2

The ZymoBIOMICS® Microbial Community Standard contains three easy-to-lyse bacteria, five tough-to-lyse bacteria, and two tough-to-lyse yeasts.

Note: The copy number of rDNA genes in bacteria can vary a lot and contribute to 16S counts appear deferentially affected depending on the bacterial host. This is well documented for the mock community used in this experiment in the Zymo protocol document [4].

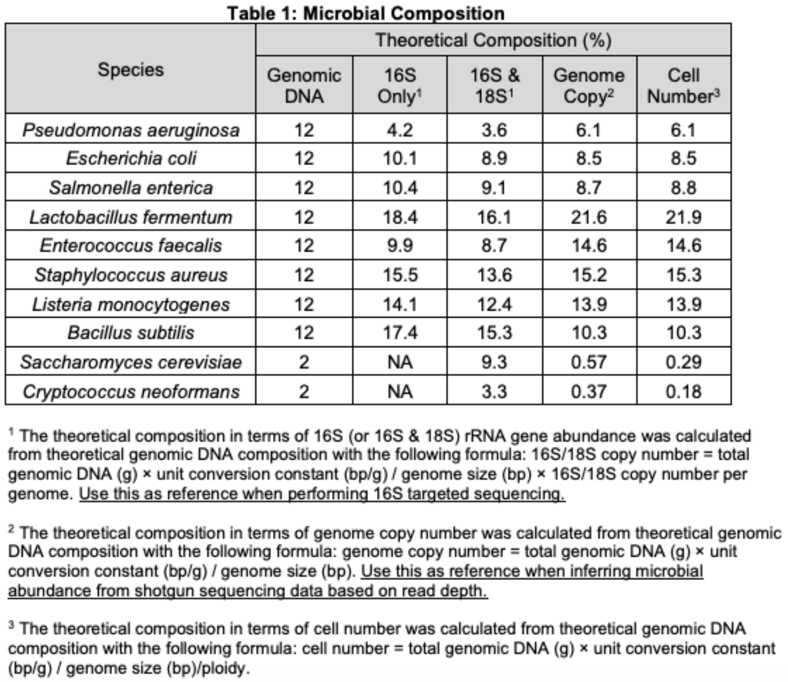


Table 1: Microbial Composition

Species	Theoretical Composition (%)				
	Genomic DNA	16S Only[1]	16S & 18S[1]	Genome Copy[2]	Cell Number[3]
Pseudomonas aeruginosa	12	4.2	3.6	6.1	6.1
Escherichia coli	12	10.1	8.9	8.5	8.5
Salmonella enterica	12	10.4	9.1	8.7	8.8
Lactobacillus fermentum	12	18.4	16.1	21.6	21.9
Enterococcus faecalis	12	9.9	8.7	14.6	14.6
Staphylococcus aureus	12	15.5	13.6	15.2	15.3
Listeria monocytogenes	12	14.1	12.4	13.9	13.9
Bacillus subtilis	12	17.4	15.3	10.3	10.3
Saccharomyces cerevisiae	2	NA	9.3	0.57	0.29
Cryptococcus neoformans	2	NA	3.3	0.37	0.18

[1] The theoretical composition in terms of 16S (or 16S & 18S) rRNA gene abundance was calculated from theoretical genomic DNA composition with the following formula: 16S/18S copy number = total genomic DNA (g) × unit conversion constant (bp/g) / genome size (bp) × 16S/18S copy number per genome. Use this as reference when performing 16S targeted sequencing.

[2] The theoretical composition in terms of genome copy number was calculated from theoretical genomic DNA composition with the following formula: genome copy number = total genomic DNA (g) × unit conversion constant (bp/g) / genome size (bp). Use this as reference when inferring microbial abundance from shotgun sequencing data based on read depth.

[3] The theoretical composition in terms of cell number was calculated from theoretical genomic DNA composition with the following formula: cell number = total genomic DNA (g) × unit conversion constant (bp/g) / genome size (bp)/ploidy.

[3] https://files.zymoresearch.com/datasheets/ds1706_zymobiomics_microbial_community_standards_data_sheet.pdf
[4] https://files.zymoresearch.com/protocols/_d6305_d6306_zymobiomics_microbial_community_dna_standard.pdf

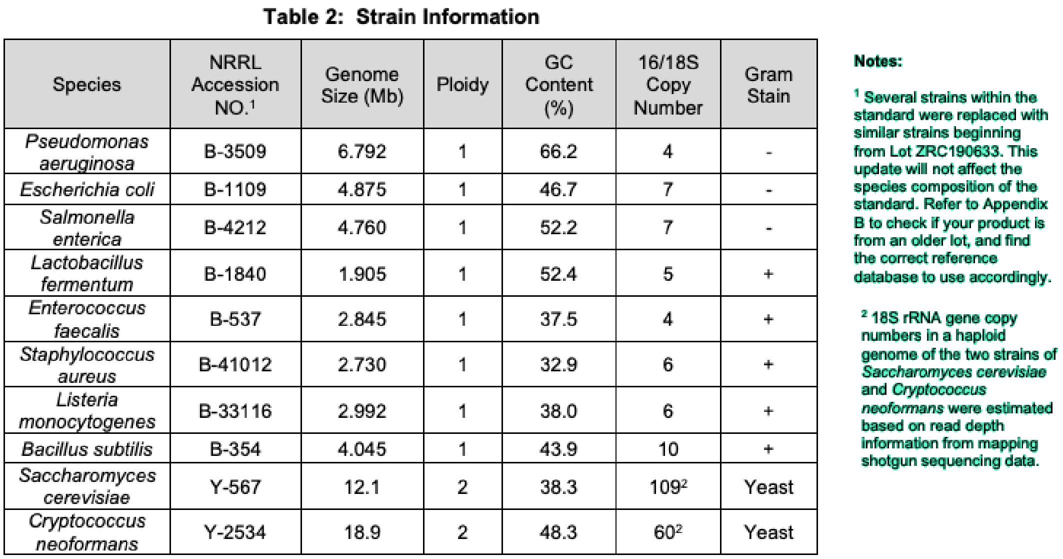


Table 2: Strain Information

Species	NRRL Accession NO.[1]	Genome Size (Mb)	Ploidy	GC Content (%)	16/18S Copy Number	Gram Stain
Pseudomonas aeruginosa	B-3509	6.792	1	66.2	4	-
Escherichia coli	B-1109	4.875	1	46.7	7	-
Salmonella enterica	B-4212	4.760	1	52.2	7	-
Lactobacillus fermentum	B-1840	1.905	1	52.4	5	+
Enterococcus faecalis	B-537	2.845	1	37.5	4	+
Staphylococcus aureus	B-41012	2.730	1	32.9	6	+
Listeria monocytogenes	B-33116	2.992	1	38.0	6	+
Bacillus subtilis	B-354	4.045	1	43.9	10	+
Saccharomyces cerevisiae	Y-567	12.1	2	38.3	109[2]	Yeast
Cryptococcus neoformans	Y-2534	18.9	2	48.3	60[2]	Yeast

Notes:

[1] Several strains within the standard were replaced with similar strains beginning from Lot ZRC190633. This update will not affect the species composition of the standard. Refer to Appendix B to check if your product is from an older lot, and find the correct reference database to use accordingly.

[2] 18S rRNA gene copy numbers in a haploid genome of the two strains of *Saccharomyces cerevisiae* and *Cryptococcus neoformans* were estimated based on read depth information from mapping shotgun sequencing data.

Note that yeast have copy numbers one order of magnitude higher than bacteria, although not relevant here it will be in the case of mixed populations between yeast and bacteria and amplicons for both 16S and 18S (or ITS).

Nextflow install and setup

The nextflow pipeline is available from github and should be cloned locally on the analysis server. After download of the github repo, the first nextflow run gets 3 Docker images and downloads the classification databases.

During the first run, the docker components (N=3) will be downloaded and installed in the default Docker cache on the server

- kpinpb/pb-16s-nf-tools
- kpinpb/pb-16s-nf-qiime
- kpinpb/pb-16s-vis

```
git clone git@github.com:PacificBiosciences/pb-16S-nf.git
cd pb-16S-nf

nextflow run main.nf \
  --download_db \
  -profile docker

# a new 'databases' folder is added to the nextflow folder
#  GTDB_bac120_arc53_ssu_r207_fullTaxo.fa.gz
#  GTDB_ssu_all_r207.qza
#  GTDB_ssu_all_r207.taxonomy.qza
#  RefSeq_16S_6-11-20_RDPv16_fullTaxo.fa.gz
#  silva-138-99-seqs.qza
#  silva-138-99-tax.qza
#  silva_nr99_v138.1_wSpecies_train_set.fa.gz
```

Nextflow test

A built-in test data can be used to validate the install as follows

```
# Create test_sample.tsv for testing
echo -e "sample-id\tabsolute-filepath\ntest_data\t$(readlink -f test_data/test_1000_reads.fastq.gz)" > test_data/test_sample.tsv

nextflow run main.nf \
  --input test_data/test_sample.tsv \
  --metadata test_data/test_metadata.tsv \
  --outdir test_results \
  -profile docker
```

The run should take only few minutes and produce a folder with intermediate data and results as discussed later in this report.

Nextflow Zymo run

Two text files need to be prepared based on the available read sets; a sample manifest and a metadata file (as standard in QIIME2):

- manifest: a run_sample.tsv file that relates the sample names (barcode-pairs) and the full path to each fastq file
- metadata: a run_metadata.tsv file that relates the same sample names to sample groups or conditions used in the wet-lab experiment (info provided by the customer)

The two files used in this test run are reproduced below

- run_sample.tsv

sample-id	absolute-file-path
4170_bc1005–bc1096	/data/analyses/Zymo-SequelIIe-Hifi/reads/4170_bc1005–bc1096.fastq.gz
4285_bc1022–bc1107	/data/analyses/Zymo-SequelIIe-Hifi/reads/4285_bc1022–bc1107.fastq.gz
4296_bc1022–bc1060	/data/analyses/Zymo-SequelIIe-Hifi/reads/4296_bc1022–bc1060.fastq.gz
4112_bc1008–bc1075	/data/analyses/Zymo-SequelIIe-Hifi/reads/4112_bc1008–bc1075.fastq.gz
4128_bc1005–bc1107	/data/analyses/Zymo-SequelIIe-Hifi/reads/4128_bc1005–bc1107.fastq.gz
4356_bc1005–bc1112	/data/analyses/Zymo-SequelIIe-Hifi/reads/4356_bc1005–bc1112.fastq.gz

Note: The path should be the full path to the fastq files, if you copy the files elsewhere, do not forget to adapt the second column to reflect your storage.

- run_metadata.tsv

sample_name	condition
4170_bc1005–bc1096	control
4285_bc1022–bc1107	control
4296_bc1022–bc1060	control
4112_bc1008–bc1075	control
4128_bc1005–bc1107	control
4356_bc1005–bc1112	control

the condition column should be of type 'categorical' (not numeric!)

Note that in a real experiment, the conditions will describe more sample groups than just 'control'

The following code was run to start the analysis:

```
# the full list of full path fastq can be obtained with:
# find fastq_folder -name "*.fastq.gz" -exec readlink -f {} \;

tooldir="/opt/biotools/pb-16S-nf"
cd ${tooldir}

# use >= 32 cpu for good performance
cpu=32
infolder=<path-to-indata>
sample_file=${infolder}/run_samples.tsv
metadata_file=${infolder}/run_metadata.tsv
outfolder=<path-to-outdata>

nextflow run main.nf \
  --input ${sample_file} \
  --metadata ${metadata_file} \
  --outdir ${outfolder} \
  --dada2_cpu ${cpu} \
  --vsearch_cpu ${cpu} \
  --cutadapt_cpu ${cpu} \
  -profile docker
```

The nextflow pipeline produces live output and stores all log files for inspection as well as trouble-shooting

The pipeline executes the following tasks using here default parameters

```
Launching `main.nf` [confident_fermat] DSL2 - revision: 6990708c9f

  Parameters set for pb-16S-nf pipeline for PacBio HiFi 16S
  =========================================================
  Number of samples in samples TSV: 6
  Filter input reads above Q: 20
  Trim primers with cutadapt: Yes
  Forward primer: AGRGTTYGATYMTGGCTCAG
  Reverse primer: AAGTCGTAACAAGGTARCY
  Minimum amplicon length filtered in DADA2: 1000
  Maximum amplicon length filtered in DADA2: 1600
  maxEE parameter for DADA2 filterAndTrim: 2
  minQ parameter for DADA2 filterAndTrim: 0
  Pooling method for DADA2 denoise process: pseudo
  Minimum number of samples required to keep any ASV: 1
  Minimum number of reads required to keep any ASV: 5
  Taxonomy sequence database for VSEARCH: /opt/biotools/pb-16S-nf/databases/GTDB_ssu_all_r207.qza
  Taxonomy annotation database for VSEARCH: /opt/biotools/pb-16S-nf/databases/GTDB_ssu_all_r207.taxonomy.qza
  Skip Naive Bayes classification: false
  SILVA database for Naive Bayes classifier: /opt/biotools/pb-16S-nf/databases/silva_nr99_v138.1_wSpecies_train_set.fa.gz
  GTDB database for Naive Bayes classifier: /opt/biotools/pb-16S-nf/databases/GTDB_bac120_arc53_ssu_r207_fullTaxo.fa.gz
  RefSeq + RDP database for Naive Bayes classifier: /opt/biotools/pb-16S-nf/databases/RefSeq_16S_6-11-20_RDPv16_fullTaxo.fa.gz
  VSEARCH maxreject: 100
  VSEARCH maxaccept: 100
  VSEARCH perc-identity: 0.97
  QIIME 2 rarefaction curve sampling depth: null
  Number of threads specified for cutadapt: 80
  Number of threads specified for DADA2: 80
  Number of threads specified for VSEARCH: 80
  Script location for HTML report generation: /opt/biotools/pb-16S-nf/scripts/visualize_biom.Rmd
  Container enabled via docker/singularity: true
  Version of Nextflow pipeline: 0.4

executor >  Local (24)
[54/6fb9cd] process > pb16S:write_log                  [100%] 1 of 1
[0a/64e613] process > pb16S:QC_fastq (3)               [100%] 6 of 6
[2b/de1b80] process > pb16S:cutadapt (6)               [100%] 6 of 6
[11/21c173] process > pb16S:QC_fastq_post_trim (6)     [100%] 6 of 6
[62/34b924] process > pb16S:collect_QC                 [100%] 1 of 1
[f6/adf720] process > pb16S:prepare_qiime2_manifest    [100%] 1 of 1
[90/056fed] process > pb16S:import_qiime2              [100%] 1 of 1
[54/2d5769] process > pb16S:demux_summarize            [100%] 1 of 1
[71/2b4460] process > pb16S:dada2_denoise                  [100%] 1 of 1
[bb/3f3b46] process > pb16S:filter_dada2                   [100%] 1 of 1
[5a/8b0eb6] process > pb16S:dada2_qc (1)                   [100%] 1 of 1
[25/3fb725] process > pb16S:qiime2_phylogeny_diversity (1) [100%] 1 of 1
[cf/e3bd62] process > pb16S:dada2_rarefaction (1)          [100%] 1 of 1
[ee/0759ee] process > pb16S:class_tax                      [100%] 1 of 1
[f9/fd76d0] process > pb16S:dada2_assignTax                [100%] 1 of 1
[ec/83b49d] process > pb16S:export_biom                    [100%] 1 of 1
[65/285e74] process > pb16S:barplot_nb (1)                 [100%] 1 of 1
[72/66021f] process > pb16S:barplot (1)                    [100%] 1 of 1
[7e/d6da7d] process > pb16S:html_rep (1)                   [100%] 1 of 1
[0d/213bfe] process > pb16S:krona_plot                     [100%] 1 of 1
Completed at: 20-Dec-2022 13:46:47
Duration    : 14m 10s
CPU hours   : 17.4
Succeeded   : 35
```

NOTE: By default, the rarefaction depth is computed automatically to include 80% of the samples. In circumstances where all samples need to be included, an arbitrary depth based on the smaller sample(s) can be given as argument in the run command `--rarefaction_depth <observed smaller depth>` and changing the name of the output folder. This can also be done after the standard run (80%) using an extra argument `-resume --rarefaction_depth <observed smaller depth>` without having to rerun the whole pipeline (will overwrite the previous results where recomputed!).

REM: The command above was rerun to take all samples into account, based on the read count slightly over 10'000 observed for the smallest sample.

```
# the full list of full path fastq can be obtained with:
# find fastq_folder -name "*.fastq.gz" -exec readlink -f {} \;

tooldir="/opt/biotools/pb-16S-nf"
cd ${tooldir}

# use >= 32 cpu for good performance
cpu=32
infolder=<path-to-indata>
sample_file=${infolder}/run_samples.tsv
metadata_file=${infolder}/run_metadata.tsv
outfolder=<path-to-outdata>

nextflow run main.nf \
  --input ${sample_file} \
  --metadata ${metadata_file} \
  --outdir ${outfolder}_10k \
  --dada2_cpu ${cpu} \
  --vsearch_cpu ${cpu} \
  --cutadapt_cpu ${cpu} \
  --rarefaction_depth 10000 \
  -profile docker
```

```
Launching `main.nf` [confident_fermat] DSL2 - revision: 6990708c9f

  Parameters set for pb-16S-nf pipeline for PacBio HiFi 16S
  =========================================================
  Number of samples in samples TSV: 6
  Filter input reads above Q: 20
  Trim primers with cutadapt: Yes
  Forward primer: AGRGTTYGATYMTGGCTCAG
  Reverse primer: AAGTCGTAACAAGGTARCY
  Minimum amplicon length filtered in DADA2: 1000
  Maximum amplicon length filtered in DADA2: 1600
  maxEE parameter for DADA2 filterAndTrim: 2
  minQ parameter for DADA2 filterAndTrim: 0
  Pooling method for DADA2 denoise process: pseudo
  Minimum number of samples required to keep any ASV: 1
  Minimum number of reads required to keep any ASV: 5
  Taxonomy sequence database for VSEARCH: /opt/biotools/pb-16S-nf/databases/GTDB_ssu_all_r207.qza
  Taxonomy annotation database for VSEARCH: /opt/biotools/pb-16S-nf/databases/GTDB_ssu_all_r207.taxonomy.qza
  Skip Naive Bayes classification: false
  SILVA database for Naive Bayes classifier: /opt/biotools/pb-16S-nf/databases/silva_nr99_v138.1_wSpecies_train_set.fa.gz
  GTDB database for Naive Bayes classifier: /opt/biotools/pb-16S-nf/databases/GTDB_bac120_arc53_ssu_r207_fullTaxo.fa.gz
  RefSeq + RDP database for Naive Bayes classifier: /opt/biotools/pb-16S-nf/databases/RefSeq_16S_6-11-20_RDPv16_fullTaxo.fa.gz
  VSEARCH maxreject: 100
  VSEARCH maxaccept: 100
  VSEARCH perc-identity: 0.97
  QIIME 2 rarefaction curve sampling depth: 10000
  Number of threads specified for cutadapt: 80
  Number of threads specified for DADA2: 80
  Number of threads specified for VSEARCH: 80
  Script location for HTML report generation: /opt/biotools/pb-16S-nf/scripts/visualize_biom.Rmd
  Container enabled via docker/singularity: true
  Version of Nextflow pipeline: 0.4

executor >  Local (35)
[16/961acc] process > pb16S:write_log                      [100%] 1 of 1
[18/96f98f] process > pb16S:QC_fastq (3)                   [100%] 6 of 6
[fd/90ad41] process > pb16S:cutadapt (6)                   [100%] 6 of 6
[87/d217c6] process > pb16S:QC_fastq_post_trim (6)         [100%] 6 of 6
[3b/1d189f] process > pb16S:collect_QC                     [100%] 1 of 1
[77/b05b00] process > pb16S:prepare_qiime2_manifest        [100%] 1 of 1
[05/7b5f13] process > pb16S:import_qiime2                  [100%] 1 of 1
[ea/578c10] process > pb16S:demux_summarize                [100%] 1 of 1
[bb/3dc94d] process > pb16S:dada2_denoise                  [100%] 1 of 1
[f3/8da1e4] process > pb16S:filter_dada2                   [100%] 1 of 1
[1d/f0b337] process > pb16S:dada2_qc (1)                   [100%] 1 of 1
```

```
[76/9c0007] process > pb16S:qiime2_phylogeny_diversity (1) [100%] 1 of 1
[2c/44c38b] process > pb16S:dada2_rarefaction (1)          [100%] 1 of 1
[02/5a6f6a] process > pb16S:class_tax                      [100%] 1 of 1
[1d/f3a1b4] process > pb16S:dada2_assignTax                [100%] 1 of 1
[4c/051df4] process > pb16S:export_biom                    [100%] 1 of 1
[14/a4333e] process > pb16S:barplot_nb (1)                 [100%] 1 of 1
[0c/f2be75] process > pb16S:barplot (1)                    [100%] 1 of 1
[86/8cbd2e] process > pb16S:html_rep (1)                   [100%] 1 of 1
[ea/771e6b] process > pb16S:krona_plot                     [100%] 1 of 1
Completed at: 20-Dec-2022 14:08:30
Duration    : 13m 45s
CPU hours   : 17.3
Succeeded   : 35
```

Zymo run results

After running all Zymo samples, the standard output can be inspected and part of it shared with the customer

The main output folder has the following standard structure:

```
cutadapt_summary
dada2
filtered_input_FASTQ
import_qiime
nb_tax
parameters.txt
results
summary_demux
trimmed_primers_FASTQ
```

The **results** folder contains symbolic links to all final key files and can be forwarded to the customer as-is

```
alpha-rarefaction-curves.qzv
best_tax_merged_freq_tax.tsv
best_taxonomy.tsv
best_taxonomy_withDB.tsv
best_tax.qza
dada2_qc.tsv
dada2_stats.qzv
dada2_table.qzv
feature-table-tax.biom
feature-table-tax_vsearch.biom
krona_html/
krona.qzv
merged_freq_tax.qzv
phylogeny_diversity/ (and subfolder: core-metrics-diversity/)
rarefaction_depth_suggested.txt
reads_QC/
samplefile.txt
stats.tsv
tax_export/
taxonomy_barplot_nb.qzv
taxonomy_barplot_vsearch.qzv
taxonomy.vsearch.qza
visualize_biom.html
vsearch_merged_freq_tax.tsv
```

The results of this run are shared next to this report in the **Zymo-SequelIIe-Hifi__results__local** folder to allow more exploration of this typical data.

All files ending with *.qzv* are QIIME2 visualization files that can be fed to the online QIIME2-Viewer (https://view.qiime2.org/) to create and customize plots or tables.

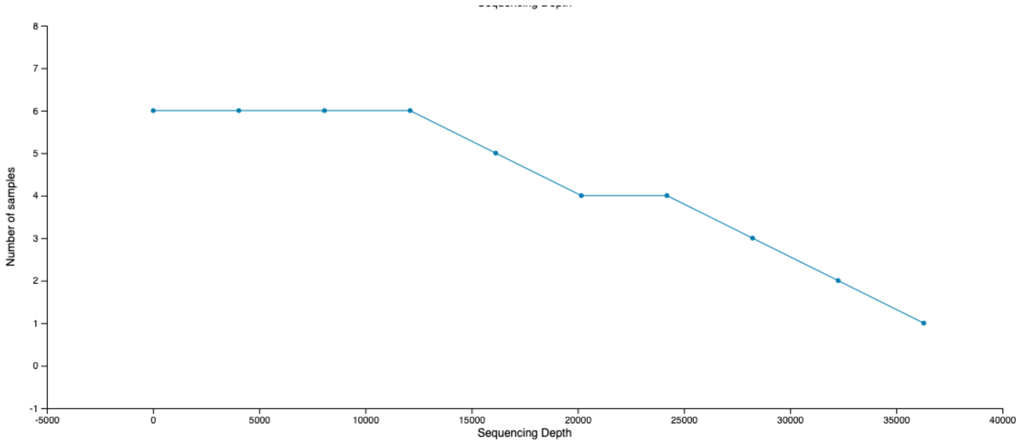
Files with extension *.qza* are QIIME2 objects that can be reloaded in QIIME2 to proceed in the analysis while files with extension *.tsv* are data files that can be used for further analysis (eg. in *R*).

The main file present in the results folder is the RMarkdown converted document **visualize__biom.html** reporting all QC metrics and key findings through the user browser. Tables in that file are live and can be filtered.

Examples of QIIME-View outputs

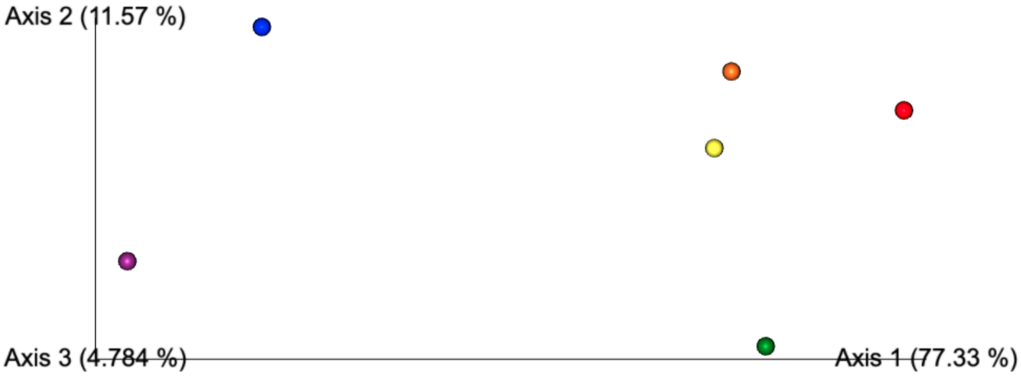
The *.qzv* files present in the results folder allow plotting using the **QIIME2 Viewer and integrated Emperor webtool** [5]

The effect of rarefaction on the number of retained samples can be seen by plotting the file **alpha-rarefaction-curves.qzv**.

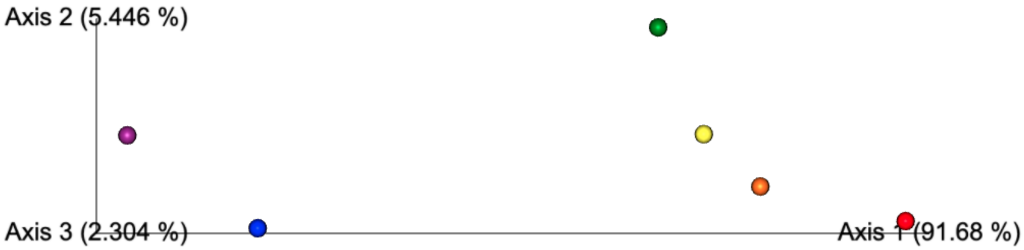


As other illustration we show below two of the plots produced by the pipeline and showing multi dimensional principal component analyses results

- Bray Curtis



- Weighted unifrac

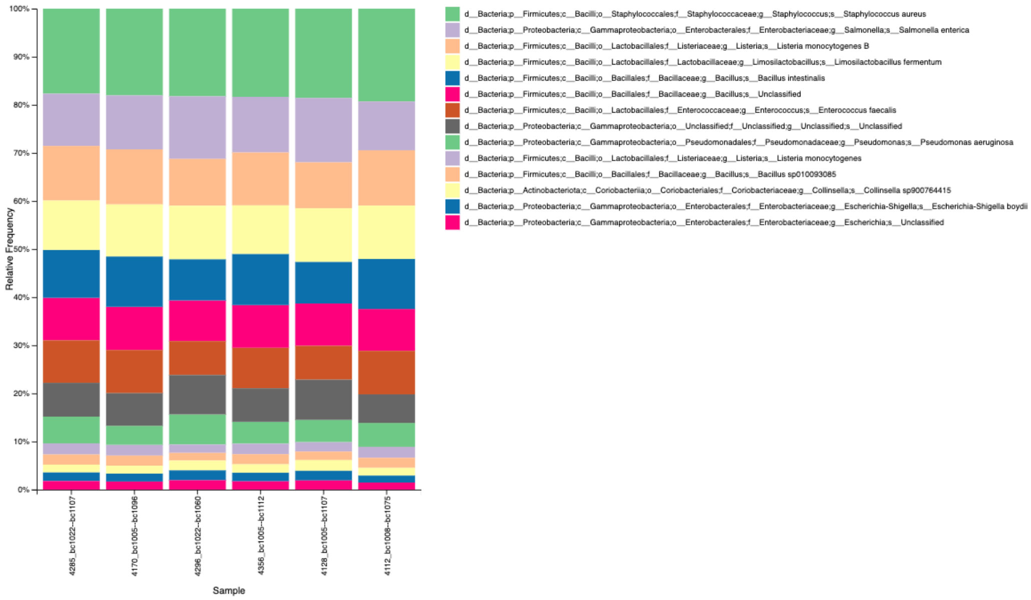


The viewer can also produce pictures for other results present in the results folder among which the two classification outcomes produced by the workflow for our 5 Zymo samples. The *Vsearch* classification is sometimes more accurate

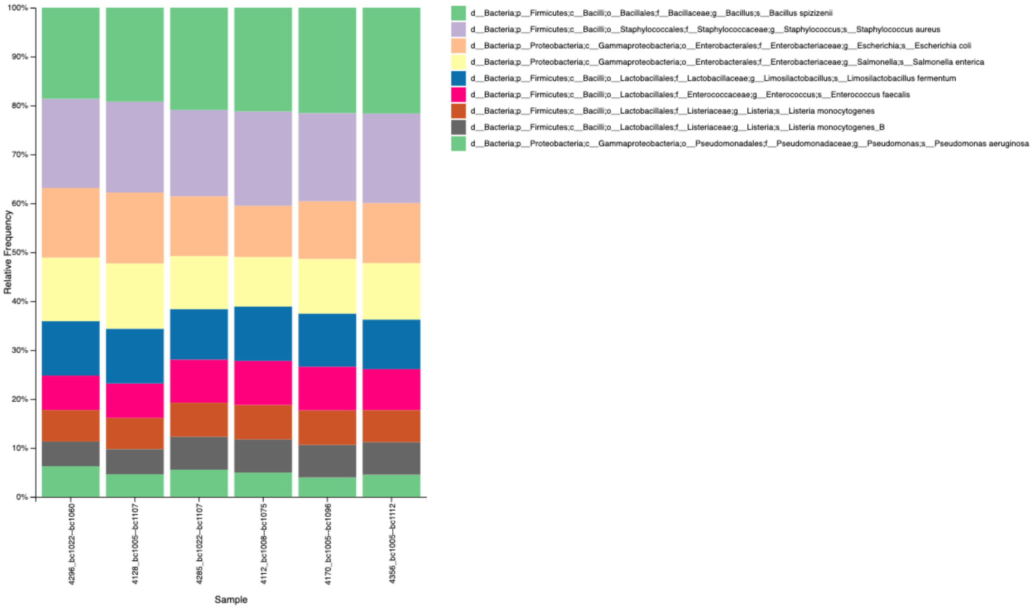
[5] https://view.qiime2.org/

and is based only on **Vsearch** best hits while the **Naive Bayes Classifier** is based on multiple search results and may be more complete but may sometimes include absent species (see doc on github)).

- taxonomy__barplot__nb.qzv

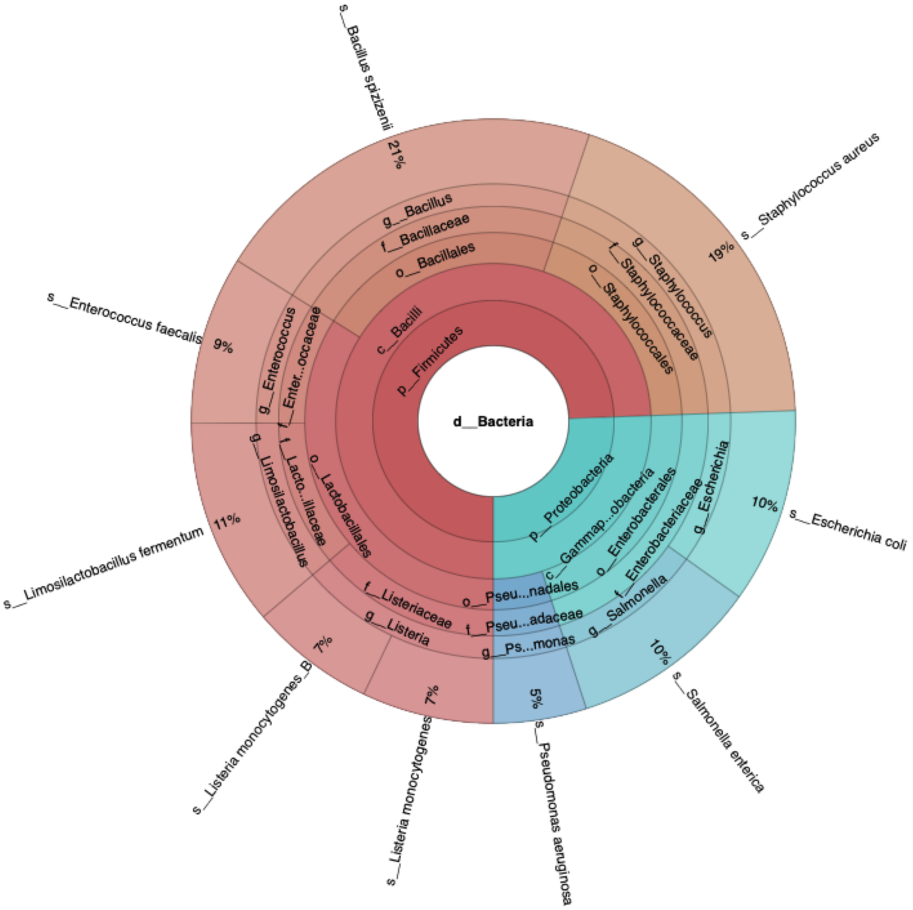


- taxonomy__barplot__vsearch.qzv



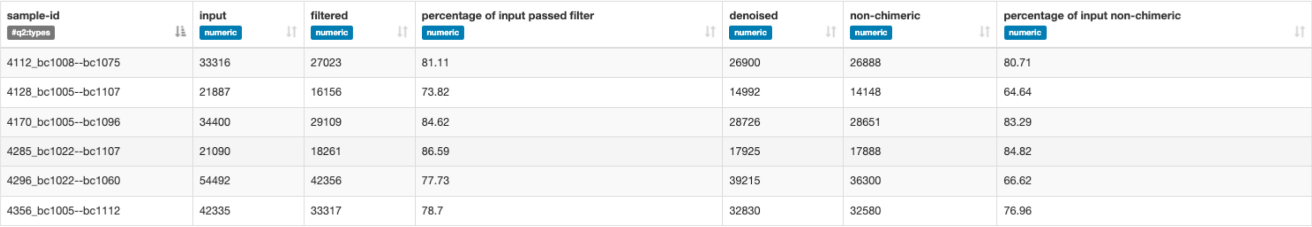
When compared to the theoretical distribution shown in the first part of this report, experimental results are nicely concordant.

Additional plots can be produced like the **Krona** classification shown next (sample bc1008-bc1075)



Various tables (.qzv) can be converted to pretty tables in the viewer as well.

As example, the content of **dada2__stats.qzv** (this table is also present in the html report)

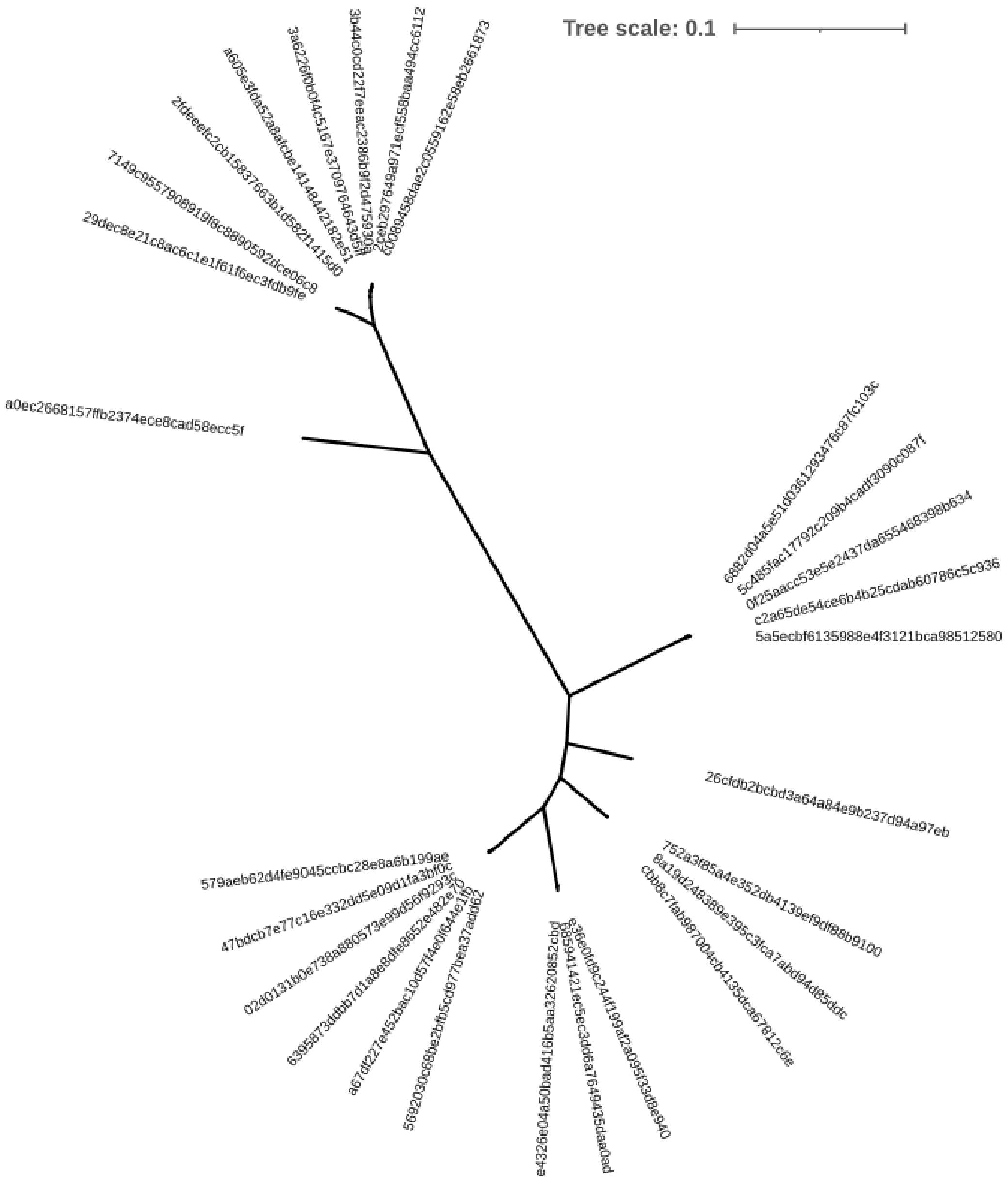


sample-id #q2:types	input numeric	filtered numeric	percentage of input passed filter numeric	denoised numeric	non-chimeric numeric	percentage of input non-chimeric numeric
4112_bc1008--bc1075	33316	27023	81.11	26900	26888	80.71
4128_bc1005--bc1107	21887	16156	73.82	14992	14148	64.64
4170_bc1005--bc1096	34400	29109	84.62	28726	28651	83.29
4285_bc1022--bc1107	21090	18261	86.59	17925	17888	84.82
4296_bc1022--bc1060	54492	42356	77.73	39215	36300	66.62
4356_bc1005--bc1112	42335	33317	78.7	32830	32580	76.96

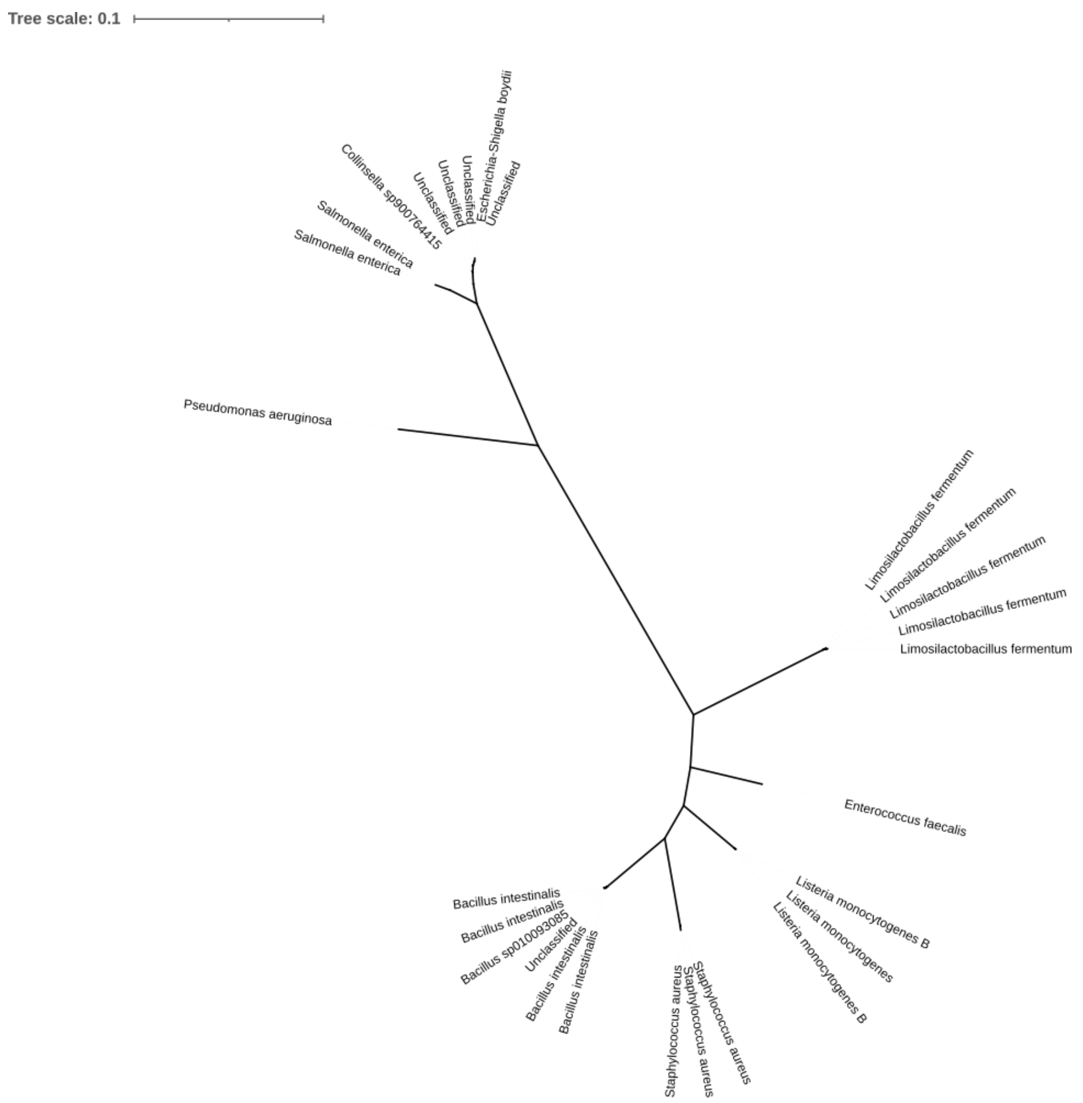
A phylogeny tree can be converted to a picture using **TreeViewer** [6] or **iTol** [7] and the file *phylotree_mafft_rooted.nwk* (note that, in the current version of the pipeline, the the tree shows the ASV labels rather than the stain names)

REM: iTol outputs are shown next

[6] https://github.com/arklumpus/TreeViewer
[7] https://itol.embl.de/


Tree scale: 0.1
29dec8e21c8ac6c1e1f61f6ec3fdb9fe
7149c9557908919f8c8890592dce06c8
2fdeeefc2cb15837663b1d582f1415d0
a605e3fda52a8afcbe14148442182e51
3a6226f0b0f4c5167e3709764643d5ff
3b44c0cd22f7eeac2386b9f2d475930a
2ceb297649a971ecf558baa494cc6112
60089458dae2c0559162e58eb2661873
a0ec2668157ffb2374ece8cad58ecc5f
6882d04a5e51d036129347 6c87fc103c
5c485fac17792c209b4cadf3090c087f
0f25aacc53e5e2437da655468398b634
c2a65de54ce6b4b25cdab60786c5c936
5a5ecbf6135988e4f3121bca98512580
26cfdb2bcbd3a64a84e9b237d94a97eb
752a3f85a4e352db4139ef9df88b9100
8a19d248389e395c3fca7abd94d85ddc
cbb8c7fab987004cb4135dca67812c6e
e36e0fd9c244f199af2a095f33d8e940
68594142 1ec5ec3dd6a7649435daa0ad
e4326e04a50bad416b5aa32620852cbd
5692030c68be2bfb5cd977bea37add62
a67df227e452bac10d57f4e0f644e1f0
6395873ddbb7d1a8e8dfe8652e482e70
02d0131b0e738a880573e99d56f9293c
47bdcb7e77c16e332dd5e09d1fa3bf0c
579aeb62d4fe9045ccbc28e8a6b199ae

A little *R* magic can replace the Feature.ID used for tip labels with genus and species extracted from the taxonomy results.


Tree scale: 0.1
Escherichia-Shigella boydii
Unclassified
Unclassified
Unclassified
Unclassified
Collinsella sp900764415
Salmonella enterica
Salmonella enterica
Pseudomonas aeruginosa
Limosilactobacillus fermentum
Limosilactobacillus fermentum
Limosilactobacillus fermentum
Limosilactobacillus fermentum
Limosilactobacillus fermentum
Enterococcus faecalis
Listeria monocytogenes B
Listeria monocytogenes
Listeria monocytogenes B
Staphylococcus aureus
Staphylococcus aureus
Staphylococcus aureus
Bacillus intestinalis
Bacillus intestinalis
Bacillus sp010093085
Unclassified
Bacillus intestinalis
Bacillus intestinalis

Conclusion

The **pb-16s-nf** pipeline is *fast* and *stable* and allows analysis of multiple barcoded samples obtained from PacBio Sequel-IIe runs with minimal user setup (preparing both *samples.tsv* and *metadata.tsv*).

Note: Additional metadata columns can be passed to the tool and will be saved in the qiime2 objects,allowing advanced graphics production and follow-up analyses.

In this particular mini-project (synthetic community), multiple ASV of the same genus+species suggest different 16S copies with different sequences present in the same bacterium present in the mock community.

The height expected species were found back, although with slightly different names as compared to the Zymo doc. Some with a unique ASV (Pseudomonas.aeruginosa and Enterococcus.faecalis) while others were represented by multiple ASVs (eg. Escherichia.Coli and Bacilus.subtilis [intestinalis] both with 6 ASV detected).

The pipeline produces a variety of outputs that can be fed to other Qiime2 methods or imported into R

For more advanced follow-up analyzes, the user can use the many tools present in the **Qiime2 package** and get support from the **Qiime2 Forum** as well as import the objects in R and create custom plots or downstream analyses there using the **qiime2R R package** (see **tutorial on the qiime2 pages**)

last edits: $ThuDec22, 2022$

VIB
NUCLEOMICS
CORE

more at **http://www.nucleomics.be**