1 Sample report

1.1 General Information

This sample has the code "mock" and is named "Mock community with 38 species". It is part of the project code "demo_project" which is described as "Demo project for the pacmill pipeline" and contains 4 other samples.

1.2 Meta-data details

All the meta-data associated with this sample is displayed below:

- **input_dir**: ~/repos/pacmill/demo_project/
- **suffix_dir**: demo_sequence_data/
- **fwd_file_name**: mock.fastq.gz
- **fwd_read_count**: 22286
- **fwd_md5**: f4a941709e0a0b906718fcf92a70c7d7
- **contact_one_name**: Lucas Sinclair
- **contact_one_mail**: lucas.sinclair@me.com
- **contact_one_role**: consultant
- **primers_pair_name**: Long 16S-ITS-23S primers
- **fwd_primer_name**: A519F
- **fwd_primer_seq**: CAGCMGCCGCGGTAA
- **rev_primer_name**: U2428R
- **rev_primer_seq**: CCRAMCTGTCTCACGACG
- **library_strategy**: AMPLICON
- **library_source**: METAGENOMIC
- **library_selection**: PCR
- **library_layout**: SINGLE
- **platform**: PACBIO
- **instrument_model**: PacBio RS II
- **organism**: synthetic metagenome
- **sra_biosample**: SAMN10319748
- **sampling_date**: 2016-11-30 00:00:00
- **output_dir**: ~/repos/pacmill/demo_project/
- **primer_mismatches**: 2
- **primer_max_dist**: 70
- **min_read_len**: 2800
- **max_read_len**: 5500
- **phred_window_size**: 30
- **phred_threshold**: 20
- **taxa_barstacks**: phylum, class
- **max_taxa**: 18

1.3 Processing

- This report and all the analysis was generated using the pacmill python pipeline.
- Documentation and source code is available at:

https://github.com/xapple/pacmill

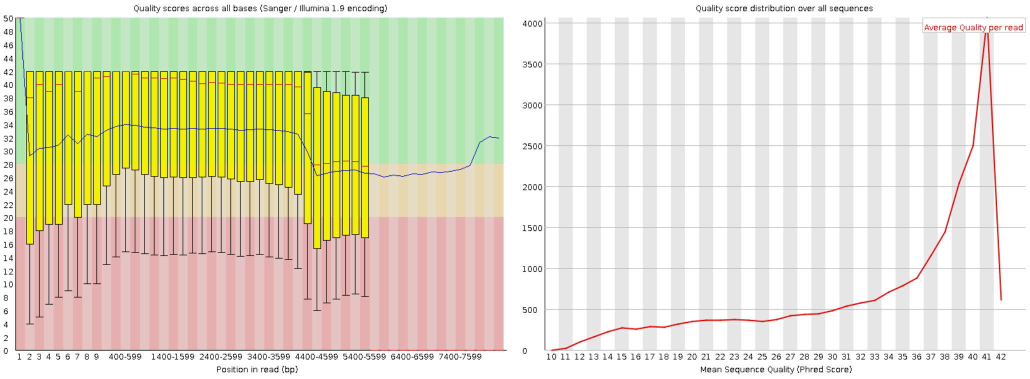
- Version `0.3.3` of the pipeline was used.
- This document was generated at `2020-08-26 04:03:09 CEST+0200` on host `bigmille`.

1.4 Validation

The original reads file weighed 51.5 MB and was validated using the https://github.com/statgen/fastQValidator program. It contained no invalid characters.

1.5 Quality

The original reads file contained 22’286 reads with an average PHRED quality of 33.37. The average quality per base and the average quality per sequence can be seen in figure 1.



(a) Per base quality

(b) Per sequence quality

Figure 1. Quality graphs from FastQC

1.6 Lengths

The shortest sequence in the file was 15 base pairs long, while the longest measured 8’307 base pairs. The complete sequence length distribution of the original reads file can be seen in figure 2.

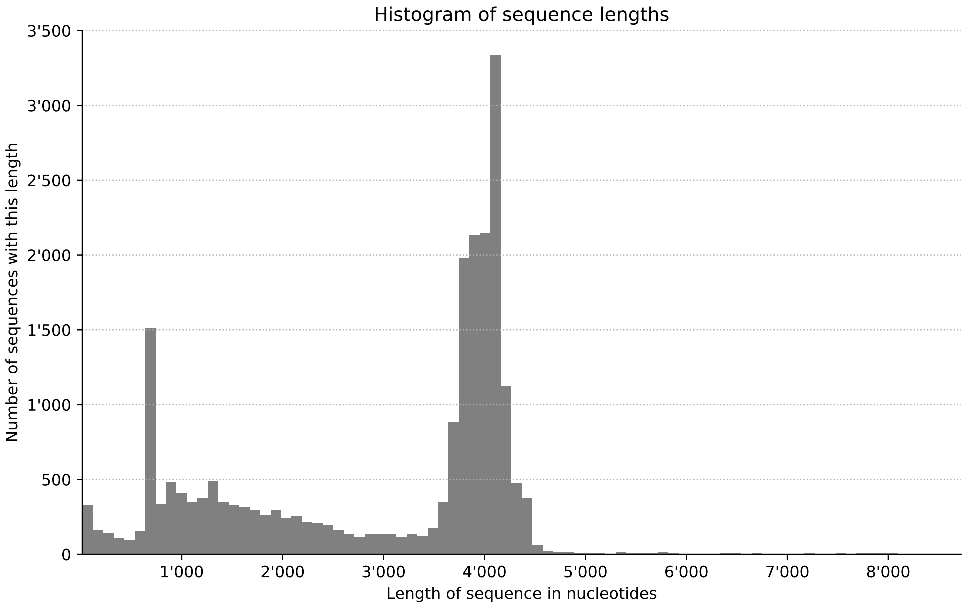


Figure 2. Distribution of sequence lengths in original reads.

1.7 Filtering

Next, we filter the sequences based on several criteria. These many sequences are lost at each step:

- Checking for the presence of both primers within at least 70 base pairs of the read's start or end and with at most 2 mismatches allowed discards 13'975 sequences (8'311 left).
- Checking for the absence of undetermined "N" bases anywhere in the reads discards 0 sequences (8'311 left).
- Checking that no sequence is shorter than 2'800 base pairs and no sequence is longer than 5'500 base pairs discards 300 sequences (8'011 left).
- Checking that no sequence has a quality score below 20 within a rolling average window of 30 base pairs discards 1'973 sequences (6'038 left).

This leaves us with 27.1% of this sample's original sequences.

1.8 Chimera Detection

```
In this step, we run every sequence through the `vsearch --uchime3_denovo` algorithm to detect the presenc

* Checking for chimeras discards 567 sequences (5'471 left).
```

1.9 Gene Presence with Barrnap

In this step, we run every sequence through the `barrnap` algorithm to detect the presence of any type of rRNA gene (via `nhmmer`). We then only keep raw sequences that had positive hits for both the 16S gene and the 23S gene. Furthermore, we extract both genes from each sequence and concatenate them together, effectively removing non-coding regions.

- Checking for the presence of rRNA genes further discards 2 sequences (5'469 left).

1.10 Clustering

Further analysis is usually performed on a collection of samples at the same time instead of on a single sample. Indeed, samples should be concatenated within a project after cleaning. The next steps are found in the project report PDF.

1.11 Taxonomy

Here is presented a summary of the taxonomic affiliation of the OTUs that were contained in this sample.

Table 1. The 20 most abundant predicted genera in this sample predicted by silva.

#	Genera	Reads
1	Pirellulaceae_unclassified	1'365
2	Caldisphaera	888
3	Puniceicoccaceae_ge	457
4	Cytophagales_unclassified	299
5	Halobacteroidaceae_unclassified	291
6	Gammaproteobacteria_unclassified	277
7	Halobacterales_unclassified	277
8	Acidimicrobiaceae_ge	260
9	Lactobacillales_unclassified	226
10	Solirubrobacteraceae_ge	171
11	Bacteria_unclassified	164
12	Azomonas	97
13	Balnearium	95
14	Lachnospiraceae_unclassified	94
15	Bacillales_unclassified	77
16	Planctomycetota_unclassified	60
17	Anaeroglobus	58
18	Leptotrichiaceae_unclassified	44
19	unknown_unclassified	37
20	Denitrovibrio	31