" Tidal and seasonal controls on cold seep activity and the efficiency of water column methanotrophs"

Authors
Tim R. de Groot [1*], Malika Menoud [2], Judith van Bleijswijk [1], Victor Hernando-Morales [1,3], Helen Czerski [4], Hossein Maazallahi[2], Sylvia Walter [2], Darci Rush [1], Thomas Röckmann[2], and Helge Niemann [1,5,6*]

Affiliations
1. NIOZ Royal Netherlands Institute for Sea Research, Texel, the Netherlands

2. Institute for Marine and Atmospheric (IMAU) Research, Utrecht University, Utrecht, the Netherlands

3. Centro de Investigación Mariña (CIM), University of Vigo, Vigo, Spain

4. Department of Mechanical Engineering, University College London, London, United Kingdom

5. Department of Earth Sciences, Utrecht University, Utrecht, the Netherlands

6. CAGE – Centre of Arctic Gas Hydrate, Environment and Climate, Department of Geosciences, UiT the Arctic University of Norway, Tromsø, Norway

[*] Corresponding author: tim.de.groot@nioz.nl, helge.niemann@nioz.nl

This file includes:

Other supplementary materials

Amplicon Analysis Report pipeline Cascabel

Supplementary materials and methods

Ocean bubble camera deployment

The size of bubbles close to the sea surface were investigated with a custom build camera in-situ [1]. For this, the camera was deployed above the bubble plume at 2 m water depth in autumn. In total, the camera was deployed five times (Table S1), 2 times at ~high tide and 3 times at ~low tide. Sea conditions were fair (no breaking waves) during the entirety of each deployment. Bubbles detected during deployment and recovery were not considered. The camera continuously recorded at 15 frames sec^{-1} with an exposure time of 20 µs. Out of ~130000 frames, 43 contained bubbles; several frames recorded multiple bubbles. The camera measurements revealed smaller bubbles (average radius = 0.4 ± 0.15 mm, n = 7) approximately during low tide and comparably larger bubbles (average. radius = 0.82 ± 0.49 mm, n = 39) approximately at high tide (Fig. S2). The average radius of rising bubbles from Darci's Site close to the sea surface was thus 0.73 ± 0.49 mm.

Endmember mixing modelling

Linear mixing of surface waters equilibrated with atmospheric CH_4 and bottom waters with maximum CH_4 concentrations was modelled based on an endmember approach. Equilibrium with atmospheric CH_4 was assumed for surface waters (15 °C summertime, 14 °C autumn) translating to CH_4 concentrations of ~ 2.6 nM [2] with a $\delta^{13}C$-value of ~ -47 ‰ and δD-value of ~ -85 ‰. The second endmember was bottom water as found at low tide with maximum CH_4 concentration in summer at 35 m water depth at 22 h (i.e., 2082 nM, $\delta^{13}C$-CH_4 = -82 ‰; δD-CH_4 = -185 ‰) and in autumn at 35 m water depth at 6 h (i.e., 496 nM, $\delta^{13}C$-CH_4 = -78 ‰; δD-CH_4 = -193 ‰).

Rayleigh fractionation modelling

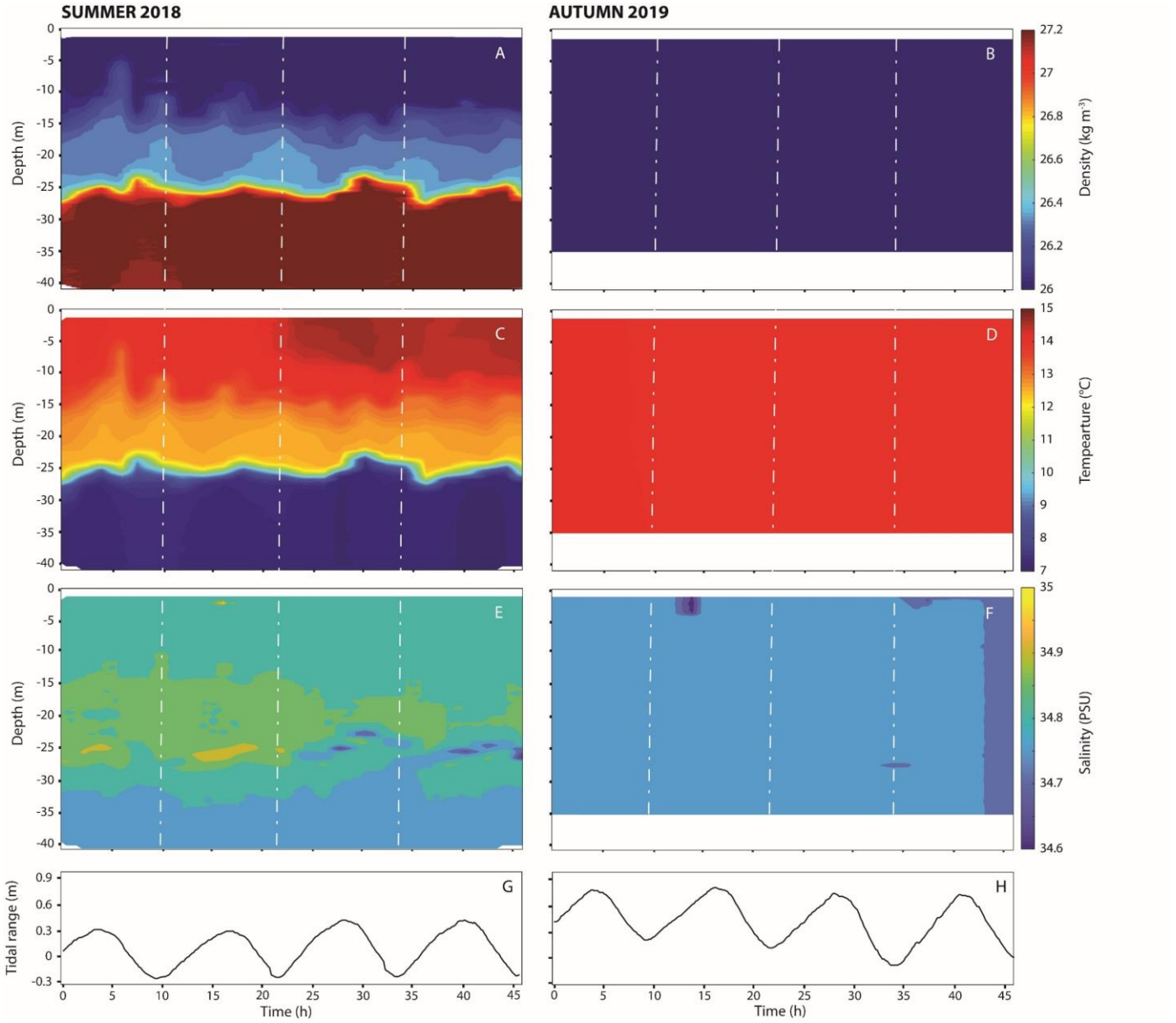
A Rayleigh fractionation model was applied for carbon and hydrogen stable isotope compositions to test whether enrichment in heavy isotopes (^{13}C, D) was related to CH_4 oxidation by methanotrophs[3]. For this, samples with CH_4 concentrations of 60 nM ($\delta^{13}C$-CH_4 ≃ -79 ‰, δD-CH_4 ≃ -184 ‰) in summer and 55 nM ($\delta^{13}C$-CH_4 ≃ -79 ‰, δD-CH_4 ≃ -189 ‰) in autumn were defined as the CH_4 source signal and thus starting point of the Rayleigh fractionation model. The apparent isotope enrichment ε is then calculated according to an open system approach[4,5]:

$$\delta_s = \delta_{s,0} + \varepsilon_{open} \ln f \qquad (6)$$

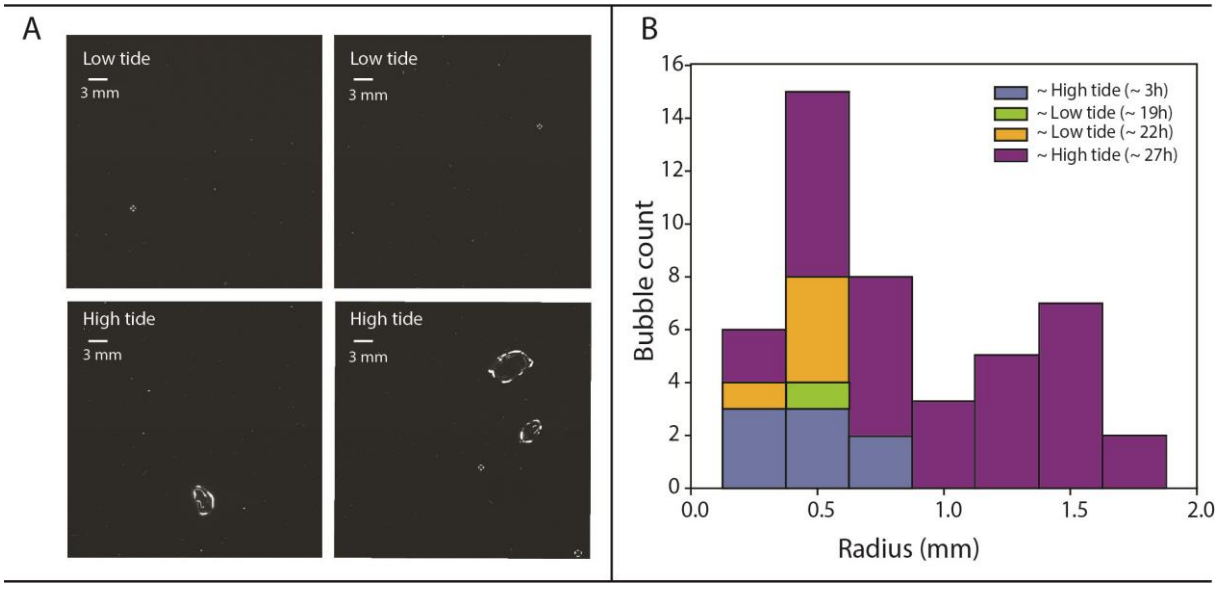
Where δ_s and $\delta_{s,0}$ are the $\delta^{13}C$ or δD values of CH_4 at a given concentration and of the source signal, respectively. *f* denotes the fraction of the remaining CH_4. Apparent ε-values for $\delta^{13}C$-CH_4 (ε^C) were -7.8 ‰ in summer and -7.6 ‰ in autumn (Fig. S5), which are consisted with previous reported ε^C values from deep sea [6,7] and estuarine [8] environments. Apparent ε-values for δD-CH_4 (ε^H) were -15.9 ‰ in summer and higher than previous reported ε^H values from the marine environment [5-7]. Note that δD was measured in few autumn samples with CH_4 concentrations < 54 nM, precluding to calculate Rayleigh distillation for H/D in a well constrained manner.

Bootstrap statistics

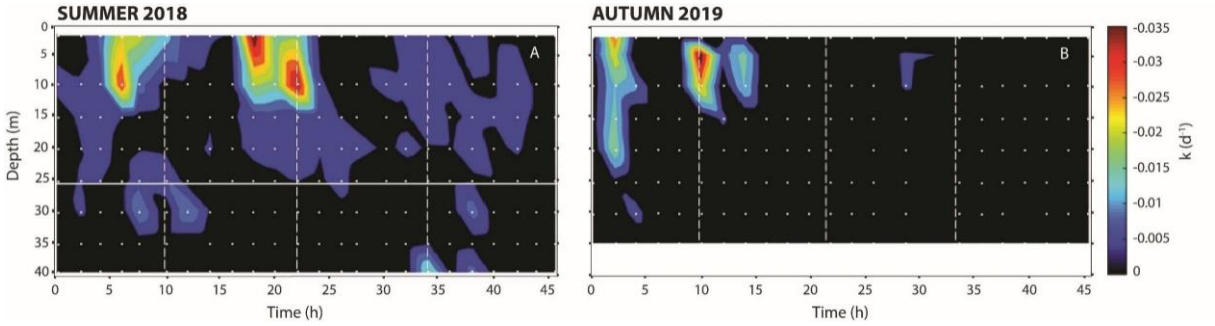
Bootstrapping (with replacement) was used to estimate the distribution of the statistic mean of the weighted average CH_4 concentration in summer (average = 137.5 ± 30.8 nM) and autumn (average = 46.7 ± 2.3 nM)[9]. Bootstrapping was conducted on the total number of samples available (Summer: n = 24; Autumn: n = 22) and on a random selection of a subset of samples (with n = 20, n = 15, n = 10, n = 3) to determine the probable effect of lower resampling (Table S2). In total, 1500 bootstraps per sample were determined (k = 1500). The Bootstrap standard error of all bootstrap samples were calculated and compared with the actual standard deviation of measured samples.



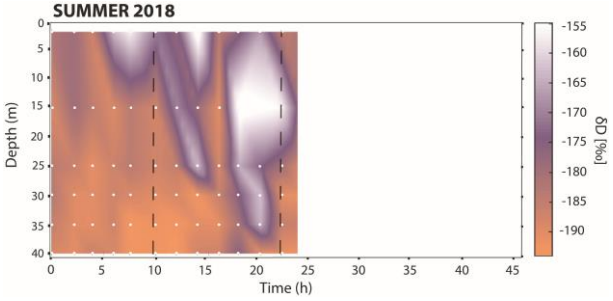
Supplementary Figure 1. Water column properties at Darci's Site. Spatiotemporal distribution of water column density (A, B), temperature (C, D) and salinity (E, F). Vertical dotted lines indicate low tide.



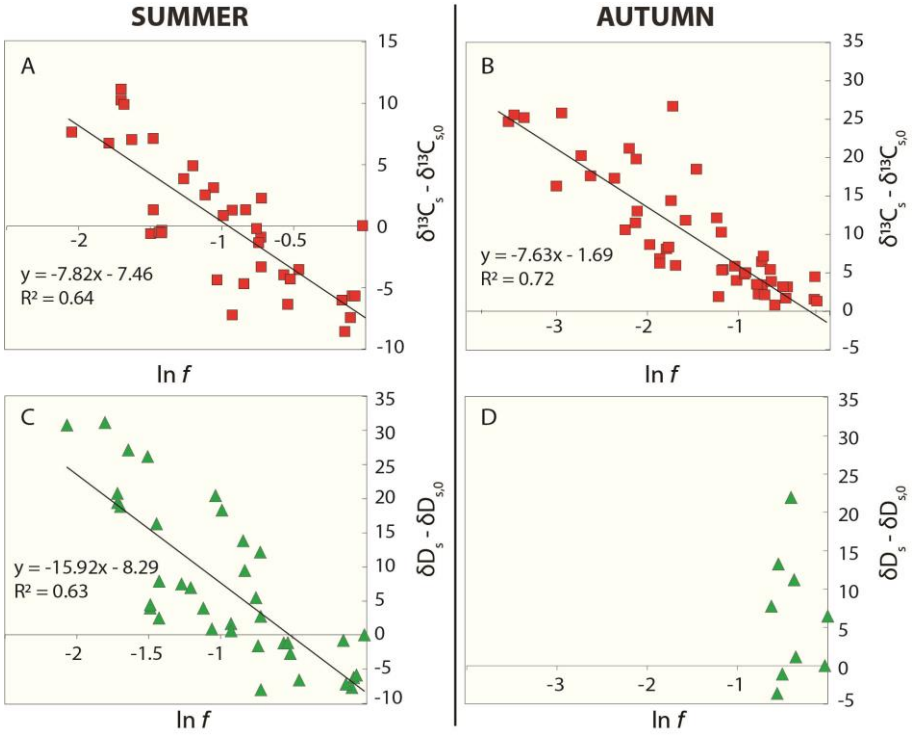
Supplementary Figure 2. In situ measurements of rising bubbles. (A) Exemplary images of bubbles captured at 2 m water depth using the high-resolution oceanic bubble camera[1]. The average radius of bubbles was 0.4 ± 0.15 mm (n = 7) during low tide and 0.82 ± 0.49 mm (n = 39) during high tide. (B) Stacked histogram depicting the distribution of bubble radii (n = 46) measured exclusively during the autumn season. The blue and purple colors indicate high tide observations after approximately 3 and 27 hours into the time-series, respectively. The green and yellow colors represent low tide observations after approximately 19 and 22 hours into the time-series, respectively.



Supplementary Figure 3. First order rate constants at Darci's Site. Spatiotemporal distribution of the first order rate constants (*k*) in summer (A) and autumn (B). White dots denote actual sampling depth/time. The pycnocline (only present in summertime) is depicted as a white horizontal line (A). Vertical pink dotted lines indicate low tide.



Supplementary Figure 4. Stable hydrogen isotopic composition. Spatiotemporal distribution of stable hydrogen isotopic composition of CH_4 in summer. **S**amples were only taken during the first 23 hours of the time-series. In autumn, only a limited number of samples were analysed for δD-CH_4 precluding spatiotemporal representation. White dots denote actual sampling depth/time. The pycnocline is depicted as a white horizontal line. Vertical pink dotted lines indicate low tide.



Supplementary Figure 5. Isotope fractionation. Rayleigh fractionation of stable isotopes of carbon (A, B) and hydrogen (C, D) in summer and autumn. Samples with CH_4 concentrations of 60 nM (δ^{13}C-CH_4 ≃ -79‰, δD-CH_4 ≃ -184‰) in summer and 55 nM (δ^{13}C-CH4 ≃ -79‰, δD-CH_4 ≃ -189‰) in autumn were defined as the CH_4 source signal ($CH_{4(0)}$, δ_0). The apparent isotope enrichment ε was then calculated according to an open system approach. Note that δD-CH_4 was measured in too few autumn samples with CH_4 concentrations < 55 nM, precluding to calculate Rayleigh distillation for H/D in a well constrained manner.

Supplementary Table 1. Ocean bubble camera deployment data in autumn. Total deployment time was 138 minutes including entering and exiting surface water. Bubbles seen during entry and exit of deployment were ignored.

Autumn 2019				
Date	**UTC Time (hour)**	**Length (min)**	**# Frames with bubbles**	**# Bubbles**
12.10.2019	14:27	26	7	8
13.10.2019	6:12	35	1	1
13.10.2019	9:18	28	5	5
13.10.2019	14:03	31	30	32
14.10.2019	6:00	26	0	0

Supplementary Table 2. Bootstrap resampling (k = 1500, alpha = 0.05) of weighted averages of CH_4 concentrations in summer (n = 24, average = 137.5 ± 30.8 nM) and autumn (n = 22, average = 46.7 ± 2.3 nM) in which n = amount of resampling. Resampling < 10 times increased the SE steeply in summer, but to a lesser extent in autumn.

Summer 2018				
n	**Average (nM)**	**SE (nM)**	**Lower bound (nM)**	**Upper bound (nM)**
3	131.3	76.6	33.4	304
10	136.9	46.4	66.5	236
15	138.5	38.2	75.0	225
20	135.8	32.0	77.0	209
24	137.0	30.7	83.0	200
Autumn 2019				
n	**Average (nM)**	**SE (nM)**	**Lower bound (nM)**	**Upper bound (nM)**
3	46.7	5.8	34.3	55.2
10	46.8	3.3	39.7	52.4
15	46.8	2.8	41.0	51.8
20	46.7	2.4	41.6	51.1
22	46.8	2.1	42.4	50.7

Supplementary Table 3. Methane concentrations and activity of MOBs at reference stations 35km NNW (55° 36.492 N, 3° 57.352 E) in summer and 176 km SSE (53° 46.338 N, 4° 46.078 E) in autumn to Darci's Site.

Summer 2018				
#R	**Depth (m)**	**[CH_4] nM**	**k (d^{-1})**	**MOx (nM d^{-1})**
1	-35	2.47	0.006	0.014
2	-30	2.09	0.01	0.02
3	-25	1.93	0.001	0.001
4	-20	1.73	0.003	0.005
5	-15	1.81	0.005	0.009
6	-10	1.58	0.005	0.007
7	-5	1.65	0.004	0.007
8	-2	1.53	0.004	0.006
	Average	1.85	0.005	0.01
	Std. dev.	0.31	0.003	0.01
	Autumn 2019			
#R	**Depth (m)**	**[CH_4] nM**	**k (d^{-1})**	**MOx (nM d^{-1})**
1	-35	3.05	0.003	0.008
2	-30	2.89	0.003	0.01
3	-25	2.9	0	0
4	-20	2.79	0.001	0.003
5	-15	2.93	0.01	0.029
6	-10	2.74	0.001	0.003
7	-5	2.93	0.002	0.006
8	-2	2.88	0	0
	Average	2.89	0.002	0.007
	Std. dev.	0.09	0.003	0.009

Supplementary Table 4. MOB primers used for qPCR. Primer Met2_338F is based on a previously published type 2 MOB primer[10].

Primer pair	Primer sequence 5' – 3'	Targets	Annealing temp (°C)
Met2_338F short Methyob_1_665R	CTACGGGAGGCAGCAG TCTCTCGGACTCAAGACTAT	Methyloceanibacter	65
OM43F OM43_765R	GAGCTCAACTTGGGAACTGCG CGTGCATGAGCGTCAGTGTTAT	OM43 clade i.e. Methylophilus methylotrophus Methylocystis rosea Methylosinus sporium	65
Mprofun445F Mprofun846R	GGGAGGAAAACAGGCAGGT TGTATATAGGCCCAACGGCT	Methyloprofundus sedimenti	68

References

1 Czerski, H. *et al.* Ocean bubbles under high wind conditions – Part 1: Bubble distribution and development. *Ocean Science* **18**, 565-586, doi:10.5194/os-18-565-2022 (2022).

2 Wiesenberg, D. A. & Guinasso, N. L. Equilibrium Solubilities of Methane, Carbon Monoxide, and Hydrogen in Water and Sea Water. *Journal of Chemical & Engineering Data* **24**, 356-360 (1979).

3 Whiticar, M. J. Carbon and hydrogen isotope systematics of bacterial formation and oxidation of methane. *Chemical Geology* **161**, 291-314, doi:10.1016/S0009-2541(99)00092-3 (1999).

4 Mariotti, A. *et al.* Experimental determination of nitrogen kinetic isotope fractionation: Some principles; illustration for the denitrification and nitrification processes. *Plant and Soil* **62**, 413-430, doi:10.1007/BF02374138 (1981).

5 Jacques, C. *et al.* Carbon and Hydrogen Isotope Signatures of Dissolved Methane in the Scheldt Estuary. *Estuaries and Coasts* **44**, 137-146, doi:10.1007/s12237-020-00768-3 (2021).

6 Kawagucci, S. *et al.* Hydrogen and carbon isotope fractionation factors of aerobic methane oxidation in deep-sea water. *Biogeosciences* **18**, 5351-5362, doi:10.5194/bg-18-5351-2021 (2021).

7 Kawagucci, S. *et al.* Gas geochemical characteristics of hydrothermal plumes at the HAKUREI and JADE vent sites, the Izena Cauldron, Okinawa Trough. *Geochemical Journal* **44**, 507-518, doi:10.2343/geochemj.1.0100 (2010).

8 Sansone, F. J., Holmes, M. E. & Popp, B. N. Methane stable isotopic ratios and concentrations as indicators of methane dynamics in estuaries. *Global Biogeochemical Cycles* **13**, 463-474, doi:10.1029/1999GB900012 (1999).

9 Roth, F. *et al.* High spatiotemporal variability of methane concentrations challenges estimates of emissions across vegetated coastal ecosystems. *Global Change Biology* **28**, 4308-4322, doi:10.1111/gcb.16177 (2022).

10 Lane D, J. 16S/23S rRNA sequencing. *Nucleic acid techniques in bacterial systematics*, 115-175 (1991).

Other supplementary materials

Amplicon Analysis Report for Library: NIOZ153

CASCABEL is designed to run amplicon sequence analysis across single or multiple read libraries.

The objective of this pipeline is to create different output files which allow the user to explore data in a simple and meaningful way, as wellas facilitate downstream analysis, based on the generated output files.

Another aim of **CASCABEL** is also to encourage the documentation process, by creating this report in order to assure data analysisreproducibility.

User description: ASV NS only

Following you can see all the steps that were taken in order to get the final results of the pipeline.

Raw Data

The raw data for this library can be found at:

- FW raw reads: NIOZ153_309_NS/samples/NIOZ153/rawdata/fw.fastq

- RV raw reads: NIOZ153_309_NS/samples/NIOZ153/rawdata/rv.fastq

Number of total reads: 21440693

Quality Control

Evaluate quality on raw reads.

Tool: [FastQC]

Version: **FastQC v0.11.9**

Command:

```
fastqc NIOZ153_309_NS/samples/NIOZ153/rawdata/fw.fastq NIOZ153_309_NS/samples/NIOZ153/rawdata/rv.fastq --extract -o NIOZ153_309_NS/samples/NIOZ153/qc/
```

You can follow the links below, in order to see the complete FastQC report:

- FastQC for sample NIOZ153_1: FQ1

- FastQC for sample NIOZ153_2: FQ2

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
484.68	1113.59	6312.46	1099.02	1103.30	2933.20	2.60	0.00

Read pairing

Align paired end reads and merge them into one single sequence in case they overlap.Tool: [PEAR]

version: **PEAR v0.9.10 [May 30, 2016] - [+bzlib +zlib]**

Command:

```
pear -f NIOZ153_309_NS/samples/NIOZ153/rawdata/fw.fastq -r NIOZ153_309_NS/samples/NIOZ153/rawdata/rv.fastq -t 100 -v 10 -j 6 -p 0.05 -o NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/peared/seqs > NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/peared/seqs.assembled.fastq
```

Output files:

- Merged reads: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/peared/seqs.assembled.fastq

- Log file: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/peared/pear.log

Number of peared reads: 21408506 = 99.85%

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
7828.56	189.37	748.28	187.53	187.57	0.00	18821.30	0.00

Peared FastQC Analysis

Check the quality of the reads after assembly.

Tool: [FastQC]

Version: **FastQC v0.11.9**

Command:

```
fastqc NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/peared/seqs.assembled.fastq --extract -o NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/peared/qc
```

Output files:

- FastQC report: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/peared/qc/seqs.assembled_fastqc.html FQ_Report

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
735.41	314.71	3832.66	300.87	304.71	18374.66	0.80	0.00

Extract barcodes

Extract the barcodes used to identify individual samples.Tool:

[QIIME] - extract_barcodes.py

Version: **extract_barcodes.py 1.9.1**

Command:

```
extract_barcodes.py -f NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/peared/seqs.assembled.fastq -c barcode_paired_stitched --bc1_len 12 --bc2_len 12 -o NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/barcodes/
```

Output files:

- Fastq file with barcodes: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/barcodes/barcodes.fastq

- Fastq file with the reads: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/barcodes/reads.fastq

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
1248.69	119.21	5756.68	91.44	104.16	18509.66	19736.90	0.00

Correct Barcodes

Try to correct the barcode from unassigned reads.

Maximum number of mismatches **2**.

Tool: Cascabel Java program

Command:

```
java -cp Scripts/BarcodeCorrector/build/classes/ barcodecorrector.BarcodeCorrector -b
NIOZ153_309_NS/metadata/sampleList_mergedBarcodes_NIOZ153.txt -f
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/barcodes/barcodes.fastq -m 2
```

Output file:

- Barcode corrected file: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/barcodes/barcodes.fastq_corrected

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
1905.93	941.49	41548.19	928.00	931.68	2257.11	2238.32	0.00

Demultiplexing

For library splitting, also known as demultiplexing, Cascabel performs several steps to assign fragments in the original as well as reverseorientation to the correct sample.

Split samples from Fastq file

Tool: [QIIME] - split_libraries_fastq.py

version: **split_libraries_fastq.py 1.9.1**

Command:

```
split_libraries_fastq.py -m NIOZ153_309_NS/metadata/sampleList_mergedBarcodes_NIOZ153.txt -i
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/barcodes/reads.fastq -o
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/splitLibs -b
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/barcodes/barcodes.fastq_corrected -q 19 -r 5 --
retain_unassigned_reads --phred_offset 33 --barcode_type 24
```

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
3075.93	784.95	5688.65	755.84	769.18	20187.21	11276.77	0.00

Retain assigned reads

Command:

```
cat NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/splitLibs/seqs.fna | grep -P -A1 "(?!>Unass)^>" | sed '/^--$/d'
> NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/splitLibs/seqs.assigned.fna
```

Create file with only unassigned reads

Command:

```
cat NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/splitLibs/seqs.fna | grep "^>Unassigned" | sed
's/>Unassigned_[0-9]* /@/g' | sed 's/ .*//' | grep -F -w -A3 -f -
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/peared/seqs.assembled.fastq | sed '/^--$/d'
>NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/splitLibs/unassigned.fastq
```

Reverse complement unassigned reads

Tool: [Vsearch]

version: **vsearch v2.8.0_linux_x86_64, 754.8GB RAM, 144 cores**

Command:

```
vsearch --fastx_revcomp NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/splitLibs/unassigned.fastq --fastqout
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/splitLibs/unassigned.reversed.fastq
```

Barcode extraction for reverse complemented, unassigned reads

Tool: [QIIME] - extract_barcodes.py

Version: **extract_barcodes.py 1.9.1**

Command:

```
extract_barcodes.py -f NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/splitLibs/unassigned.reversed.fastq -c
barcode_paired_stitched --bc1_len 12 --bc2_len 12 -o
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/barcodes_unassigned/
```

Correct reverse complemented barcodes

Maximum number of mismatches **2**.

Tool: Cascabel Java program

Command:

```
java -cp Scripts/BarcodeCorrector/build/classes/ barcodecorrector.BarcodeCorrector -b
NIOZ153_309_NS/metadata/sampleList_mergedBarcodes_NIOZ153.txt -f
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/barcodes_unassigned/barcodes.fastq_corrected -m 2
```

Output file:

- Barcode corrected file: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/barcodes/barcodes.fastq_corrected

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
1558.47	1107.38	41548.19	1104.86	1104.93	1894.18	1893.95	0.00

Split reverse complemented reads

Tool: [QIIME] - extract_barcodes.py

Version: **extract_barcodes.py 1.9.1**

Command:

```
split_libraries_fastq.py -m NIOZ153_309_NS/metadata/sampleList_mergedBarcodes_NIOZ153.txt -i
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/barcodes_unassigned/reads.fastq -o
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/splitLibsRC -b
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/barcodes_unassigned/barcodes.fastq_corrected -q 19 -r 5 --
phred_offset 33 --barcode_type 24
```

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
3075.93	784.95	5688.65	755.84	769.18	20187.21	11276.77	0.00

Output files:

- Text histogram with the length of the fw reads: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/splitLibs/histograms.txt

- Log file for the fw reads: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/splitLibs/split_library_log.txt

- Text histogram with the length of the rv reads: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/splitLibsRC/histograms.txt

- Log file for the rv reads: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/splitLibsRC/split_library_log.txt

- Fasta file with unassigned reads: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/splitLibsRC/seqs.unassigned.fnaNumber of reads assigned on FW: 2982106 = 13.93% of the peared reads

Number of reads assigned on RVC: 3139697 = 14.67% of the peared reads

Generate single sample fastq files

Create single fastq files per samples (based on the raw data without applying any filtering).Tool:

Cascabel Java program

Command:

```
java -cp Scripts DemultiplexQiime --fasta -d
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/seqs_fw_rev_accepted.fna -o
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/demultiplexed/ -r1
NIOZ153_309_NS/samples/NIOZ153/rawdata/fw.fastq.gz -r2 NIOZ153_309_NS/samples/NIOZ153/rawdata/fw.fastq.gz
```

The demultiplexed files are located at:

- demultiplexed directory: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/demultiplexed/

- summary file: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/demultiplexed/summary.txt

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
8403.06	2135.11	41871.38	2123.66	2127.50	1845.46	1126.36	0.00

Combine reads

Concatenate forward and reverse reads.

Command:

```
cat NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/splitLibs/seqs.assigned.fna
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/splitLibsRC/seqs.assigned.fna >
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/seqs_fw_rev_accepted.fna
```

Output files:

- Fasta file with combined reads: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/seqs_fw_rev_accepted.fna

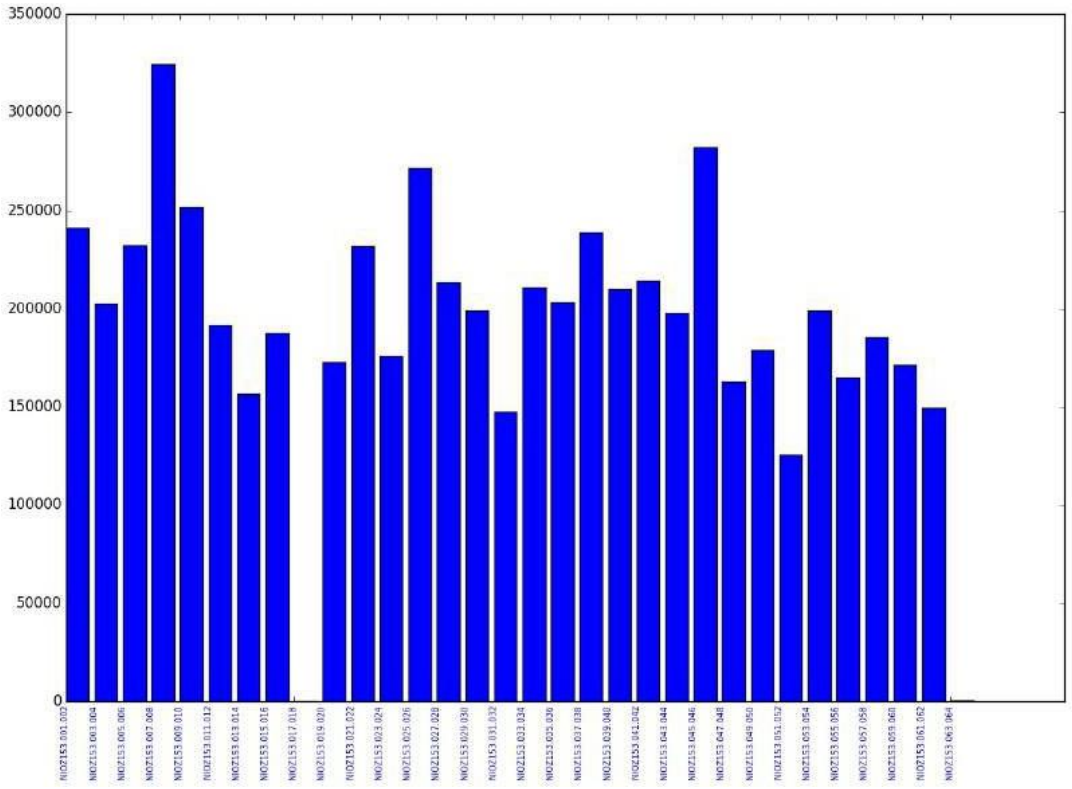
- Total number of acepted reads: 6121803 = 28.60% of the peared reads or 28.55% of the raw reads.

Benchmark data not found: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/combine_seqs_fw_rev.benchmark

Sample distribution

Sample	Seqs	prc.	Sample	Seqs	prc.	Sample	Seqs	prc.
NIOZ153.001.002	241001	3.94	NIOZ153.017.018	101	0.00	NIOZ153.033.034	210997	3.45
NIOZ153.003.004	202743	3.31	NIOZ153.019.020	172798	2.82	NIOZ153.035.036	203400	3.32
NIOZ153.005.006	232369	3.80	NIOZ153.021.022	231893	3.79	NIOZ153.037.038	238863	3.90
NIOZ153.007.008	324465	5.30	NIOZ153.023.024	175824	2.87	NIOZ153.039.040	210218	3.43
NIOZ153.009.010	251591	4.11	NIOZ153.025.026	271478	4.43	NIOZ153.041.042	214513	3.50
NIOZ153.011.012	191732	3.13	NIOZ153.027.028	213608	3.49	NIOZ153.043.044	197719	3.23
NIOZ153.013.014	156443	2.56	NIOZ153.029.030	198914	3.25	NIOZ153.045.046	282295	4.61
NIOZ153.015.016	187566	3.06	NIOZ153.031.032	147487	2.41	NIOZ153.047.048	162746	2.66

Sample	Seqs	prc.
NIOZ153.049.050	179016	2.92
NIOZ153.051.052	125490	2.05
NIOZ153.053.054	199017	3.25
NIOZ153.055.056	164970	2.69
NIOZ153.057.058	185304	3.03
NIOZ153.059.060	171437	2.80
NIOZ153.061.062	149692	2.45
NIOZ153.063.064	225	0.00



The previous chart shows the number of clean reads per sample. The bars are sorted from left to right, according to the metadata inputfile.

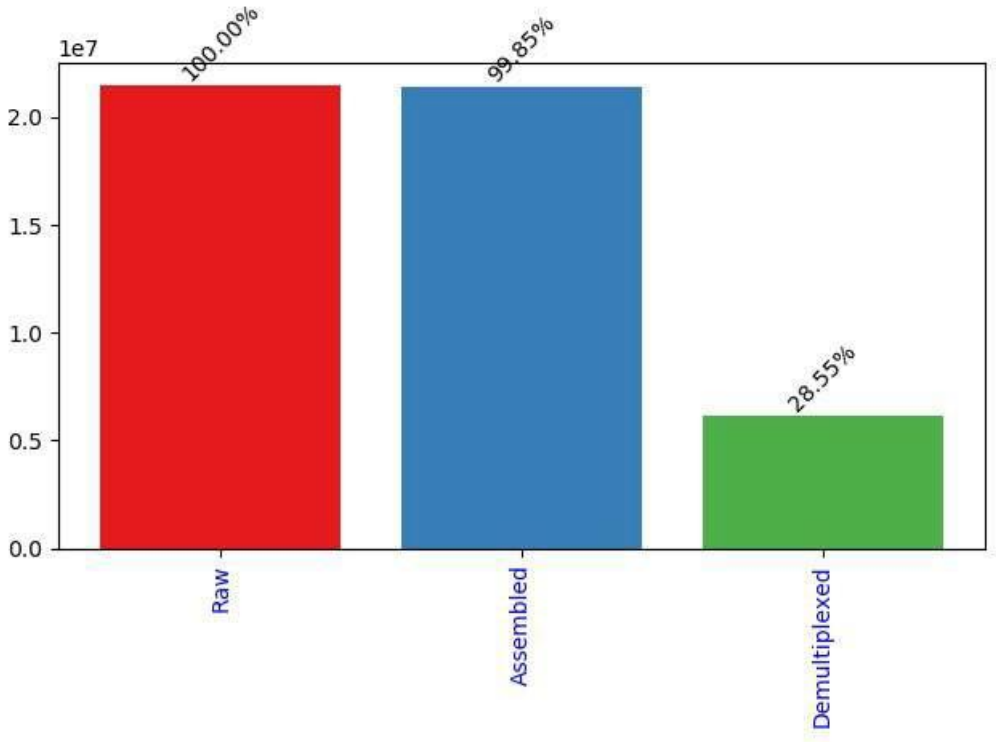
To see more details about the number of reads per sample in this library, please refer to the file:

NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/seqs_fw_rev_filtered.dist.txt

Final counts

Following you can see the final read counts:

File description	Location	Number of reads	Prc(%) vs raw
Raw reads	NIOZ153_309_NS/samples/NIOZ153/rawdata/*.fq	21440693.0	100.00%
Assembled reads	NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/peared/seqs.assembled.fastq	21408506.0	99.85%
Demultiplexed reads	NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ153_data/seqs_fw_rev_accepted.fna	6121803	28.55%



ASV report

Cascabel report on downstream analyses in combination with multiple libraries (if supplied), can be found at the following link: asv_report (NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv_report_dada2.html)

References

[FastQC] *(1, 2)* FastQC v0.11.3. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data [PEAR]

PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Zhang et al (2014) Bioinformatics 30(5): 614-620 | doi:10.1093/bioinformatics/btt593

[QIIME] *(1, 2, 3, 4)* QIIME. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Gonzalez PenaA, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7(5): 335-336.

[Cutadapt] Cutadapt v1.15 .Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal, 17(1):10-12, May 2011. http://dx.doi.org/10.14806/ej.17.1.200

[Vsearch] Rognes T, Flouri T, Nichols B, Quince C, Mahé F. (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ4:e2584. doi: 10.7717/peerj.2584

Author: J. Engelmann & A. Abdala | 2021-10-28

Amplicon Analysis Report for Library: NIOZ309

CASCABEL is designed to run amplicon sequence analysis across single or multiple read libraries.

The objective of this pipeline is to create different output files which allow the user to explore data in a simple and meaningful way, as wellas facilitate downstream analysis, based on the generated output files.

Another aim of **CASCABEL** is also to encourage the documentation process, by creating this report in order to assure data analysisreproducibility.

User description: ASV NS only

Following you can see all the steps that were taken in order to get the final results of the pipeline.

Raw Data

The raw data for this library can be found at:

- FW raw reads: NIOZ153_309_NS/samples/NIOZ309/rawdata/fw.fastq

- RV raw reads: NIOZ153_309_NS/samples/NIOZ309/rawdata/rv.fastq

Number of total reads: 11472493

Quality Control

Evaluate quality on raw reads.

Tool: [FastQC]

Version: **FastQC v0.11.9**

Command:

```
fastqc NIOZ153_309_NS/samples/NIOZ309/rawdata/fw.fastq NIOZ153_309_NS/samples/NIOZ309/rawdata/rv.fastq --extract -o NIOZ153_309_NS/samples/NIOZ309/qc/
```

You can follow the links below, in order to see the complete FastQC report:

- FastQC for sample NIOZ309_1: FQ1

- FastQC for sample NIOZ309_2: FQ2

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
268.42	1109.88	6312.46	1095.84	1100.11	4235.52	2.54	0.00

Read pairing

Align paired end reads and merge them into one single sequence in case they overlap.Tool: [PEAR]

version: **PEAR v0.9.10 [May 30, 2016] - [+bzlib +zlib]**

Command:

```
pear -f NIOZ153_309_NS/samples/NIOZ309/rawdata/fw.fastq -r NIOZ153_309_NS/samples/NIOZ309/rawdata/rv.fastq -t 100 -v 10 -j 6 -p 0.05 -o NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/peared/seqs > NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/peared/seqs.assembled.fastq
```

Output files:

- Merged reads: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/peared/seqs.assembled.fastq

- Log file: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/peared/pear.log

Number of peared reads: 11440559 = 99.72%

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
4674.94	189.28	748.28	187.44	187.48	0.00	10188.18	0.00

Peared FastQC Analysis

Check the quality of the reads after assembly.

Tool: [FastQC]

Version: **FastQC v0.11.9**

Command:

```
fastqc NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/peared/seqs.assembled.fastq --extract -o
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/peared/qc
```

Output files:

- FastQC report: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/peared/qc/seqs.assembled_fastqc.html FQ_Report

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
422.38	318.14	3832.66	304.29	310.04	10124.66	0.55	0.00

Extract barcodes

Extract the barcodes used to identify individual samples.Tool:

[QIIME] - extract_barcodes.py

Version: **extract_barcodes.py 1.9.1**

Command:

```
extract_barcodes.py -f NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/peared/seqs.assembled.fastq -c
barcode_paired_stitched --bc1_len 12 --bc2_len 12 -o NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/barcodes/
```

Output files:

- Fastq file with barcodes: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/barcodes/barcodes.fastq

- Fastq file with the reads: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/barcodes/reads.fastq

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
670.04	119.19	5613.56	91.42	104.16	9989.66	10741.51	0.00

Correct Barcodes

Try to correct the barcode from unassigned reads.

Maximum number of mismatches **2**.

Tool: Cascabel Java program

Command:

```
java -cp Scripts/BarcodeCorrector/build/classes/ barcodecorrector.BarcodeCorrector -b
NIOZ153_309_NS/metadata/sampleList_mergedBarcodes_NIOZ309.txt -f
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/barcodes/barcodes.fastq -m 2
```

Output file:

- Barcode corrected file: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/barcodes/barcodes.fastq_corrected

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
1055.70	761.51	41548.19	748.03	751.71	1327.53	1314.21	0.00

Demultiplexing

For library splitting, also known as demultiplexing, Cascabel performs several steps to assign fragments in the original as well as reverseorientation to the correct sample.

Split samples from Fastq file

Tool: [QIIME] - split_libraries_fastq.py

version: **split_libraries_fastq.py 1.9.1**

Command:

```
split_libraries_fastq.py -m NIOZ153_309_NS/metadata/sampleList_mergedBarcodes_NIOZ309.txt -i
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/barcodes/reads.fastq -o
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/splitLibs -b
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/barcodes/barcodes.fastq_corrected -q 19 -r 5 --
retain_unassigned_reads --phred_offset 33 --barcode_type 24
```

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
1455.66	479.26	5747.11	451.41	464.17	11071.07	6204.91	0.00

Retain assigned reads

Command:

```
cat NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/splitLibs/seqs.fna | grep -P -A1 "(?!>Unass)^>" | sed '/^--$/d'
> NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/splitLibs/seqs.assigned.fna
```

Create file with only unassigned reads

Command:

```
cat NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/splitLibs/seqs.fna | grep "^>Unassigned" | sed
's/>Unassigned_[0-9]* /@/g' | sed 's/ .*//' | grep -F -w -A3 -f -
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/peared/seqs.assembled.fastq | sed '/^--$/d'
>NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/splitLibs/unassigned.fastq
```

Reverse complement unassigned reads

Tool: [Vsearch]

version: **vsearch v2.8.0_linux_x86_64, 754.8GB RAM, 144 cores**

Command:

```
vsearch --fastx_revcomp NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/splitLibs/unassigned.fastq --fastqout
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/splitLibs/unassigned.reversed.fastq
```

Barcode extraction for reverse complemented, unassigned reads

Tool: [QIIME] - extract_barcodes.py

Version: **extract_barcodes.py 1.9.1**

Command:

```
extract_barcodes.py -f NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/splitLibs/unassigned.reversed.fastq -c barcode_paired_stitched --bc1_len 12 --bc2_len 12 -o NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/barcodes_unassigned/
```

Correct reverse complemented barcodes

Maximum number of mismatches **2**.

Tool: Cascabel Java program

Command:

```
java -cp Scripts/BarcodeCorrector/build/classes/ barcodecorrector.BarcodeCorrector -b NIOZ153_309_NS/metadata/sampleList_mergedBarcodes_NIOZ309.txt -f NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/barcodes_unassigned/barcodes.fastq_corrected -m 2
```

Output file:

- Barcode corrected file: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/barcodes/barcodes.fastq_corrected

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
854.53	860.88	41548.19	858.37	858.43	1049.66	1029.13	0.00

Split reverse complemented reads

Tool: [QIIME] - extract_barcodes.py

Version: **extract_barcodes.py 1.9.1**

Command:

```
split_libraries_fastq.py -m NIOZ153_309_NS/metadata/sampleList_mergedBarcodes_NIOZ309.txt -i NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/barcodes_unassigned/reads.fastq -o NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/splitLibsRC -b NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/barcodes_unassigned/barcodes.fastq_corrected -q 19 -r 5 --phred_offset 33 --barcode_type 24
```

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
1455.66	479.26	5747.11	451.41	464.17	11071.07	6204.91	0.00

Output files:

- Text histogram with the length of the fw reads: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/splitLibs/histograms.txt

- Log file for the fw reads: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/splitLibs/split_library_log.txt

- Text histogram with the length of the rv reads: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/splitLibsRC/histograms.txt

- Log file for the rv reads: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/splitLibsRC/split_library_log.txt

- Fasta file with unassigned reads: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/splitLibsRC/seqs.unassigned.fnaNumber of reads assigned on FW: 2313212 = 20.22% of the peared reads

Number of reads assigned on RVC: 2319536 = 20.27% of the peared reads

Generate single sample fastq files

Create single fastq files per samples (based on the raw data without applying any filtering).Tool:

Cascabel Java program

Command:

```
java -cp Scripts DemultiplexQiime --fasta -d
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/seqs_fw_rev_accepted.fna -o
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/demultiplexed/ -r1
NIOZ153_309_NS/samples/NIOZ309/rawdata/fw.fastq.gz -r2 NIOZ153_309_NS/samples/NIOZ309/rawdata/fw.fastq.gz
```

The demultiplexed files are located at:

- demultiplexed directory: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/demultiplexed/

- summary file: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/demultiplexed/summary.txt

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
3699.77	1776.30	41866.23	1765.07	1768.92	1300.75	771.09	0.00

Combine reads

Concatenate forward and reverse reads.

Command:

```
cat NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/splitLibs/seqs.assigned.fna
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/splitLibsRC/seqs.assigned.fna >
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/seqs_fw_rev_accepted.fna
```

Output files:

- Fasta file with combined reads: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/seqs_fw_rev_accepted.fna

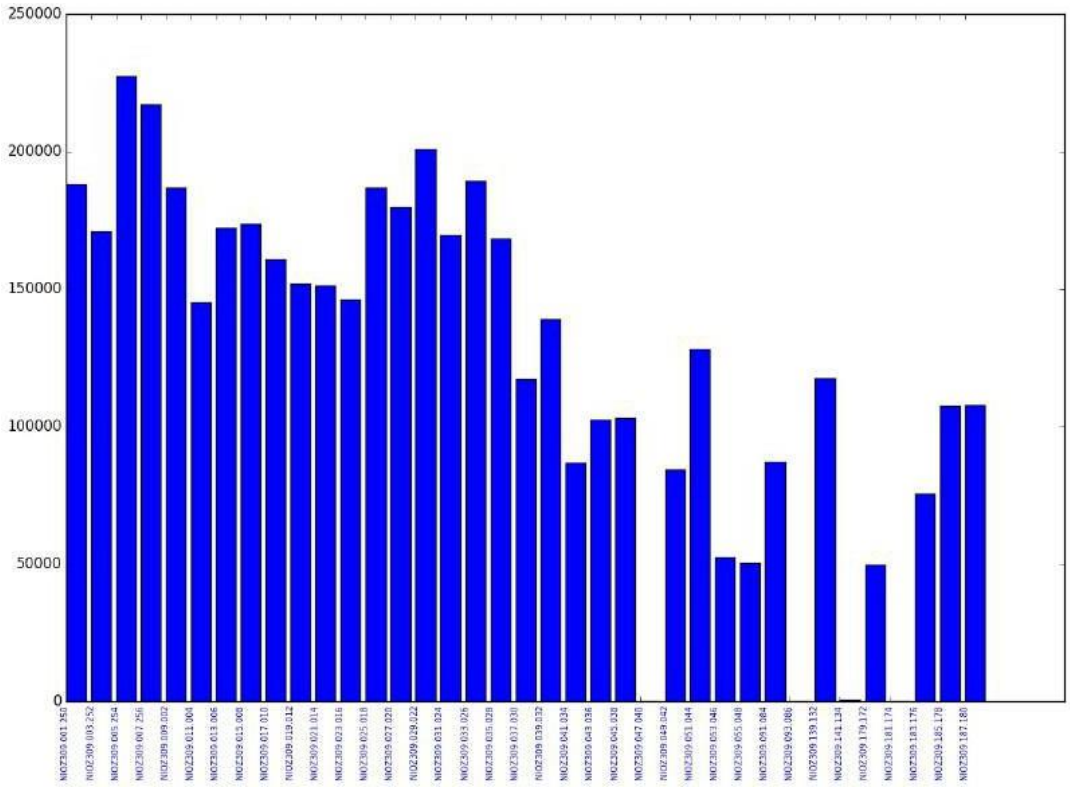
- Total number of acepted reads: 4632748 = 40.49% of the peared reads or 40.38% of the raw reads.

Benchmark data not found: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/combine_seqs_fw_rev.benchmark

Sample distribution

Sample	Seqs	prc.	Sample	Seqs	prc.	Sample	Seqs	prc.
NIOZ309.001.250	188091	4.06	NIOZ309.021.014	151105	3.26	NIOZ309.041.034	86749	1.87
NIOZ309.003.252	170829	3.69	NIOZ309.023.016	146102	3.15	NIOZ309.043.036	102638	2.22
NIOZ309.005.254	227456	4.91	NIOZ309.025.018	186806	4.03	NIOZ309.045.038	103196	2.23
NIOZ309.007.256	217336	4.69	NIOZ309.027.020	179679	3.88	NIOZ309.047.040	48	0.00
NIOZ309.009.002	186764	4.03	NIOZ309.029.022	200932	4.34	NIOZ309.049.042	84410	1.82
NIOZ309.011.004	145296	3.14	NIOZ309.031.024	169795	3.67	NIOZ309.051.044	128139	2.77
NIOZ309.013.006	172266	3.72	NIOZ309.033.026	189509	4.09	NIOZ309.053.046	52492	1.13
NIOZ309.015.008	173900	3.75	NIOZ309.035.028	168282	3.63	NIOZ309.055.048	50191	1.08
NIOZ309.017.010	160905	3.47	NIOZ309.037.030	117072	2.53	NIOZ309.091.084	87042	1.88
NIOZ309.019.012	151766	3.28	NIOZ309.039.032	139002	3.00	NIOZ309.093.086	31	0.00

Sample	Seqs	prc.
NIOZ309.139.132	117363	2.53
NIOZ309.141.134	533	0.01
NIOZ309.179.172	49610	1.07
NIOZ309.181.174	142	0.00
NIOZ309.183.176	75647	1.63
NIOZ309.185.178	107472	2.32
NIOZ309.187.180	107855	2.33



The previous chart shows the number of clean reads per sample. The bars are sorted from left to right, according to the metadata inputfile.

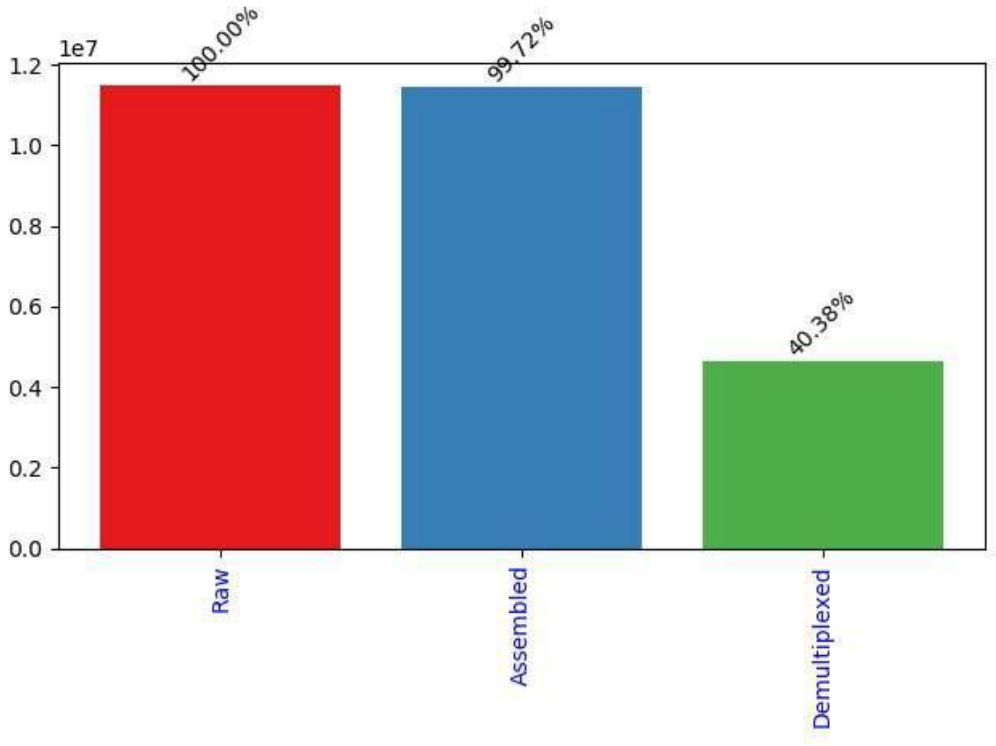
To see more details about the number of reads per sample in this library, please refer to the file:

NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/seqs_fw_rev_filtered.dist.txt

Final counts

Following you can see the final read counts:

File description	Location	Number of reads	Prc(%) vs raw
Raw reads	NIOZ153_309_NS/samples/NIOZ309/rawdata/*.fq	11472493.0	100.00%
Assembled reads	NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/peared/seqs.assembled.fastq	11440559.0	99.72%
Demultiplexed reads	NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/NIOZ309_data/seqs_fw_rev_accepted.fna	4632748	40.38%



ASV report

Cascabel report on downstream analyses in combination with multiple libraries (if supplied), can be found at the following link: asv_report (NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv_report_dada2.html)

References

[FastQC] *(1, 2)* FastQC v0.11.3. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data [PEAR]

PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Zhang et al (2014) Bioinformatics 30(5): 614-620 | doi:10.1093/bioinformatics/btt593

[QIIME] *(1, 2, 3, 4)* QIIME. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Gonzalez PenaA, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7(5): 335-336.

[Cutadapt] Cutadapt v1.15 .Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal, 17(1):10-12, May 2011. http://dx.doi.org/10.14806/ej.17.1.200

[Vsearch] Rognes T, Flouri T, Nichols B, Quince C, Mahé F. (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ4:e2584. doi: 10.7717/peerj.2584

Author: J. Engelmann & A. Abdala | 2021-10-28

Amplicon Analysis Report

CASCABEL is designed to run amplicon sequence analysis across single or multiple read libraries. This report consists of the ASV table creation and taxonomic assignment for all the combined accepted reads of given samples or libraries, if multiple.

User description: ASV NS only

Filter and Trim

Once that all the individual libraries were demultiplexed, the fastq files from all the samples for all the libraries were processed together.

The filter and trimming steps were both performed with the **filterAndTrim()** function from the R package dada2, according to user parameters.

Tool: dada2 Version:

[1] '1.19.1'

Function: filterAndTrim()

Max Expected Errors (maxEE) FW: 3

Max Expected Errors (maxEE) RV: 5

Forward read truncation: 260

Reverse read truncation: 200

Command:

```
Scripts/asvFilter.R $PWD T 260 200 3 5 10 truncQ=2, rm.phix=TRUE
NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/filter_summary.out
```

Output file:

- Filtered fastq files: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/<Library>/demultiplexed/filtered/

- Summary: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/filter_summary.out

Note: To speed up downstream computation, consider tightening maxEE. If too few reads are passing the filter, consider relaxing maxEE,perhaps especially on the reverse reads.

Make sure that your forward and reverse reads overlap after length truncation.

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
710.91	18107.14	23741.17	16052.44	16289.64	3019.73	1568.32	0.00

Amplicon Sequence Variants

In order to identify ASVs, dada2 workflow require to execute several steps. Following a summary of these steps and its main parameters. Tool: dada2

Version: **[1] '1.19.1'**

Learn errors

The first step after filtering the reads is to learn the errors from the fastq files.Function:

learnErrors(filteredFQ)

Error plots:

- FW reads error plot:: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/fw_err.pdf

- RV reads error plot:: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/rv_err.pdf

ASV inference

The amplicon sequence variant identification consists of a high resolution sample inference from the amplicon data using the learned errors.

Function: dada(filteredFQ, errors, pool='pseudo')

Merge pairs

In this step, forward and reverse reads are paired in order to create full denoised sequences.Function:

mergePairs(dadaF, dadaR)

Min overlap: 10

Max mismatch: 0

Length filtering

Sequences that are much longer or shorter than expected may be the result of non-specific priming.

- Shortest length: 362

- Longest length: 382

Remove chimeras

Sequence variants identified as bimeric are removed, and a bimera-free collection of unique sequences is generated.Function:

removeBimeraDenovo()

Method: consensus

Output files:

- Representative ASV sequences: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/representative_seq_set.fasta The total number of different ASVs is: **8348**

Assign taxonomy

Given a set of sequences, assign the taxonomy of each sequence.Tool: RDP

Function: assignTaxonomy() *implementation of RDP Classifier within dada2*

Reference database: /export/data01/databases/silva/r138.1/dada2/silva_nr99_v138.1_train_set.fa.gz The percentage of successfully assigned ASVs is: **99.98%**

Output file:

- ASV taxonomy assignation: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/representative_seq_set_tax_assignments.txt

The previous steps were performed within a Cascabel R script according to the following command:

Command

```
Scripts/asvDada2.R $PWD pseudo 10 T selfConsist=FALSE NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/ 260 220 10 T /export/data01/databases/silva/r138.1/dada2/silva_nr99_v138.1_train_set.fa.gz
```

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
93277.83	15840.02	26120.88	15827.41	15829.43	1613.25	4.69	0.00

Make ASV table

Tabulates the number of times an ASV is found in each sample, and adds the taxonomic predictions for each ASV in the last column.

Command:

```
cat NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/representative_seq_set_tax_assignments.txt | awk 'NR==FNR{if(NR>1){tax=$2;for(i=3;i<=NF;i++){tax=tax";"$i};h[$1]=tax;}next;} {if(FNR==1){print $0"\ttaxonomy"}else{print $0"\t"h[$1]}'

- NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/asv_table.txt > NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/asvTable.txt
```

Output file:

- ASV table: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/asvTable.txt

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
0.05	1.36	110.64	0.22	0.22	0.65	0.00	0.00

Convert ASV table

Convert from txt to the BIOM table format.Tool:

[BIOM]

Version: **biom, version 2.1.6**

Command:

```
biom convert -i NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/asvTable.txt -o NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/asvTable.biom --table-type 'OTU table' --table type "OTU table" --to-hdf5 --process-obs-metdata taxonomy
```

Output file:

- Biom format table: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/asvTable.biom

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
1.96	104.32	5576.21	102.11	102.14	53.40	0.20	0.00

Summarize Taxa

Summarize information of the representation of taxonomic groups within each sample. Tool:

[QIIME] - summarize_taxa.py

Version: **summarize_taxa.py 1.9.1**

Command:

```
summarize_taxa.py -i NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/otuTable.biom --level 2,3,4,5,6,7 -o NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/summary/
```

Output file:

- Taxonomy summarized counts at different taxonomy levels: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/summary/otuTable_L**N**.txt

Where **N** is the taxonomy level. Default configuration produces levels from 2 to 6.

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
22.84	154.08	5631.31	151.80	151.83	0.00	1.23	0.00

Filter ASV table

Filter ASVs from an ASV table based on their observed counts or identifier.Tool:

[QIIME] - filter_otus_from_otu_table.py

Version: **filter_otus_from_otu_table.py 1.9.1**

Minimum observation counts: 2

Command:

```
filter_otus_from_otu_table.py -i NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/asvTable.biom -o NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/asvTable_noSingletons.biom -n 2
```

Output file:

- Biom table: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/otuTable_noSingletons.biom

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
2.75	152.70	5575.36	150.38	150.41	41.88	0.21	0.00

Convert Filtered ASV table

Convert the filtered OTU table from the BIOM table format to a human readable formatTool:

[BIOM]

Version: **biom, version 2.1.6**

Command:

```
biom convert -i NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/otu/taxonomy_dada2/asvTable_noSingletons.biom -o NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/asvTable_noSingletons.txt --table-type 'OTU table' --header-key taxonomy --to-tsv
```

Output file:

- TSV format table: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/asvTable_noSingletons.txt

Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
1.97	97.99	4813.88	95.78	95.80	2.21	0.01	0.00

Filter representative sequences

Remove sequences according to the filtered OTU biom table.Tool:

[QIIME] - filter_fasta.py

Version: **filter_fasta.py 1.9.1**

Command:

```
filter_fasta.py -f NIOZ153_309_NS/samples/NorthSea_ASV_no_primer/asv/representative_seq_set.fasta -o NIOZ153_309_NS/samples/NorthSea_ASV_no_primer/asv/taxonomy_dada2/representative_seq_set_noSingletons.fasta -b NIOZ153_309_NS/samples/NorthSea_ASV_no_primer/asv/taxonomy_dada2/otuTable_noSingletons.biom
```

Output file:

- Filtered fasta file: NIOZ153_309_NS/samples/NorthSea_ASV_no_primer/asv/taxonomy_dada2/representative_seq_set_noSingletons.fasta

Krona report

Krona allows hierarchical data to be explored with zooming, multi-layered pie charts.Tool: [Krona]

These charts were created using the ASV table **without** singletonsThe report

was executed for all the samples.

Each sample is represented on a separated chart (same html report).You can see

the report at the following link:

- Krona report: kreport

Or access the html file at:

- Krona html file: NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/krona_report.html

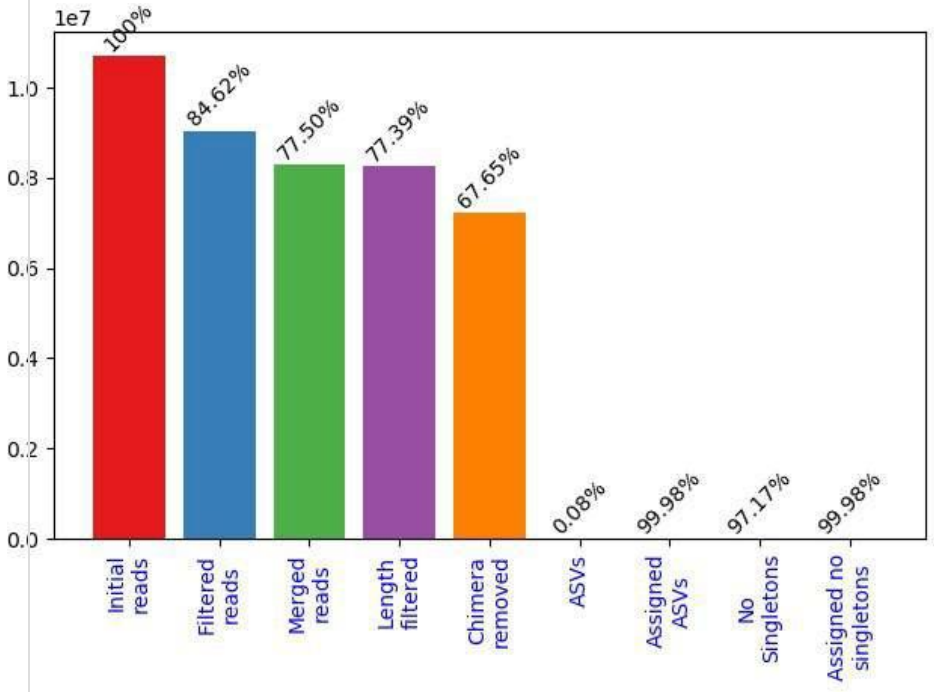
Benchmark info:

s	max_rss	max_vms	max_uss	max_pss	io_in	io_out	mean_load
4.30	33.05	387.96	26.22	27.02	10.48	9.54	0.00

Final counts

Following the read counts:

File description	Location	#	(%)
Demultiplexed reads	NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/<SAMPLE>_data/demultiplexed/*.fastq.gz	10692366	100
QA filtered & trimmed reads	NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/<LIBRARY>_data/demultiplexed/filtered/*.fastq.gz	9048193	84.6
Denoised FW reads	*No intermediate file generated*	8890950	83.1
Denoised RV reads	*No intermediate file generated*	8933928	83.5
Merged and full denoised reads	*No intermediate file generated*	8286480	77.5
Length filtered	*No intermediate file generated*	8275266	77.3
Chimera removed	*No intermediate file generated*	7233427	67.6
ASV table	NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/asvTable.txt	8348	0.08
Taxonomy assignation	NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/representative_seq_set_tax_assignments.txt	8346	99.9
ASV table (no singletons: a > 2)	NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/asv/taxonomy_dada2/asvTable_noSingletons.txt	8112	97.1
Assigned no singletons	NIOZ153_309_NS/runs/NorthSea_ASV_no_primer/otu/taxonomy_vsearch/asvTable_noSingletons.txt	8110	99.9



Note:

- Assigned ASVs percentage is the amount of successfully assigned ASVs.

- No singletons percentage is the percentage of no singletons ASVs in reference to the complete ASV table.

- Assigned No singletons is the amount of successfully no singletons assigned ASVs.

[QIIME] *(1, 2, 3)* QIIME. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Gonzalez Pena A, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010.

QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7(5): 335-336.

[Cutadapt] Cutadapt v1.15 .Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads.

EMBnet.Journal, 17(1):10-12, May 2011. http://dx.doi.org/10.14806/ej.17.1.200

[vsearch] Rognes T, Flouri T, Nichols B, Quince C, Mahé F. (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584. doi: 10.7717/peerj.2584

[Krona] Ondov BD, Bergman NH, and Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics.

2011 Sep 30; 12(1):385.

[BIOM] *(1, 2)* The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. Daniel McDonald, Jose C. Clemente, Justin Kuczynski, Jai Ram Rideout, Jesse Stombaugh, Doug Wendel, Andreas Wilke, Susan Huse, John Hufnagle, Folker Meyer, Rob Knight, and J. Gregory Caporaso.GigaScience 2012, 1:7. doi:10.1186/2047-217X-1-7

[RDP] Wang, Q, G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol. 73(16):5261-7.

[dada2] Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods, 13, 581-583. doi: 10.1038/nmeth.3869.