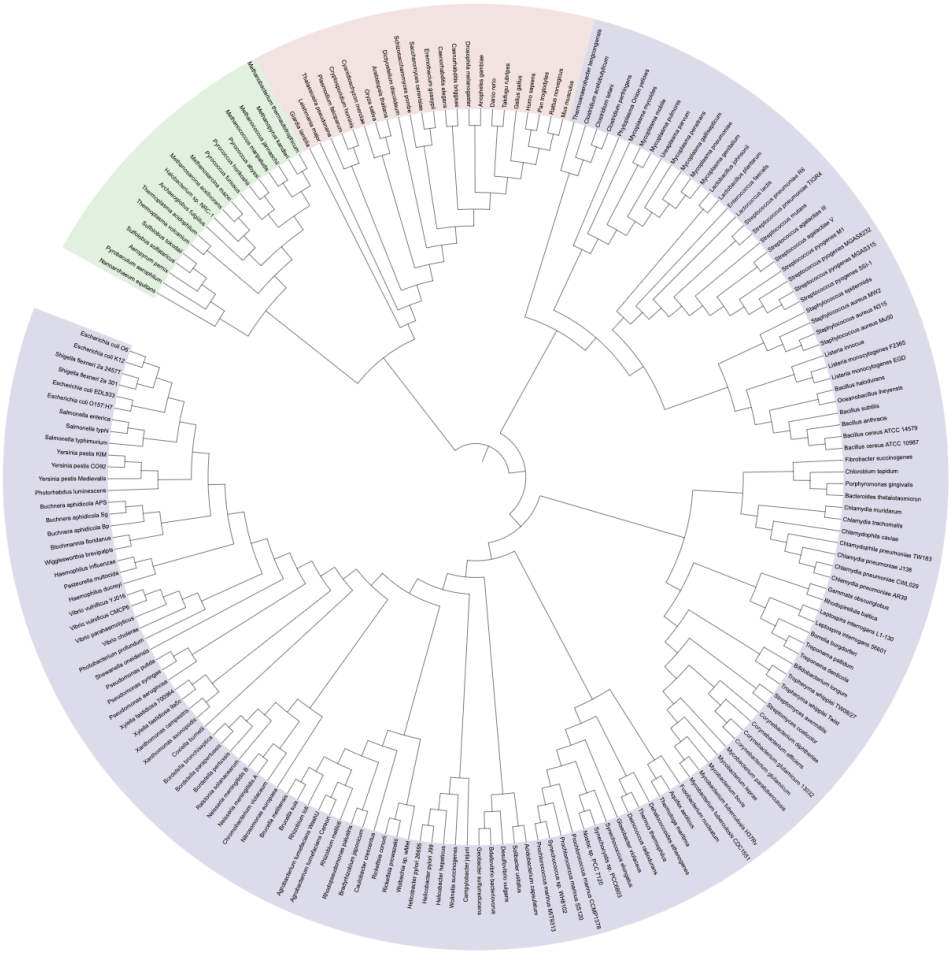
Marine Metagenomes: Diversity and Phylogeny Analyses

Practical Report 1

Year 2025–2026



Master 2 GENIOMHE

Université d'Evry Paris-Saclay

Student: Silvia Trottet, Svetlana Sannikova

Course: Metagenomics and Multiomics for microbiome studies

1 Introduction

This practical session introduces methods for analyzing microbial diversity from metagenomic sequencing data obtained by the Tara Oceans project. The main objective was to characterize and compare the microbial diversity between surface and mesopelagic water samples collected from two oceans - the North Atlantic and the Indian Ocean.

Starting from raw sequencing reads, we extracted 16S/18S rRNA gene fragments, widely used as phylogenetic markers, to assess the composition and diversity of marine microbial communities. Through a series of steps including read quality control, filtering, and clustering into Operational Taxonomic Units (OTUs), we computed several alpha- and beta-diversity indices to evaluate species richness and community similarity across environments. Finally, a phylogenetic analysis based on taxonomic annotation allowed us to identify the main taxa present in the samples.

This practical session is part of a better understanding of the diversity and structure of marine microbial communities in relation to environmental conditions such as depth, oceanic region, light availability, and nutrient concentration.

2 Results of the analysis

We will create a new conda environment for this session and install Mothur.

```
(base) svetlana@DESKTOP-LAL2IVC:~$ conda activate tp_tara
(tp_tara) svetlana@DESKTOP-LAL2IVC:~$ mothur
Linux version

Using Boost,HDF5,GSL
mothur v.1.48.0
Last updated: 5/20/22
by
Patrick D. Schloss

Department of Microbiology & Immunology

University of Michigan
http://www.mothur.org
```

Extraction of 16S/18S rRNA reads

The first step consists of extracting 16s/18s ribosomal DNA sequences, which are used very commonly for phylogenetic analyzes, from the raw read data. For this purpose, we will use the software SortMeRNA.

```
(tp_tara) svetlana@DESKTOP-LAL2IVC:~/Metagenomique$ sortmerna --ref sortmerna_ssu_db.fasta --reads reads.fastq --workdir sortmerna_out --aligned ./aligned --fastx
-e 0.001
[process:1393] === Options processing starts ... ===
Found value: sortmerna
Found flag: --ref
Found value: sortmerna_ssu_db.fasta of previous flag: --ref
Found flag: --reads
Found value: reads.fastq of previous flag: --reads
Found flag: --workdir
Found value: sortmerna_out of previous flag: --workdir
Found flag: --aligned
Found value: ./aligned of previous flag: --aligned
Found flag: --fastx
Previous flag: --fastx is Boolean. Setting to True
Found flag: -e
Found value: 0.001 of previous flag: -e
[opt_workdir:995] Using WORKDIR: "/home/svetlana/Metagenomique/sortmerna_out" as specified
[process:1483] Processing option: aligned with value: ./aligned
[process:1483] Processing option: e with value: 0.001
[process:1483] Processing option: fastx with value:
[process:1483] Processing option: reads with value: reads.fastq
[opt_reads:98] Processing reads file [1] out of total [1] files
[process:1483] Processing option: ref with value: sortmerna_ssu_db.fasta
[opt_ref:158] Processing reference [1] out of total [1] references
[opt_ref:206] File "/home/svetlana/Metagenomique/sortmerna_ssu_db.fasta" exists and is readable
[process:1503] === Options processing done ===
[process:1504] Alignment type: [best:1 num_alignments:1 min_lis:2 seeds:2]
[validate_kvdbdir:1248] Key-value DB location "/home/svetlana/Metagenomique/sortmerna_out/kvdb"
```

We can see first what quality and quantity our reads are.

```
(tp_tara) svetlana@DESKTOP-LAL2IVC:~/Metagenomique$ head reads.fastq
@ERR599021.226 PHOSPHORE:151:D0EFNACXX:2:1101:7772:2071/1
GCACACGCACACGCACACGCACACGCACACGCACACACACACGCACACGCACACGCACACGCACACGCAGACGCAGACGCACACGCAGACGCAGA
+
CCCFFFADHHHHHJJJJJHIJJJJJJJJJJJJJJJHHHFFEBDDDDD=B?=B?@@8@BB57B758;5805<9@D&08>B>5<B9>B-9<<@>&5?9<B39
@ERR599021.930 PHOSPHORE:151:D0EFNACXX:2:1101:6892:2426/1
AGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAATTTCTTCAAGAGTTTGTCCCCAGAATTTCTCCCCAGAATTTCTCCACGG
+
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJIJJJJJJJJIJJJJJIIJJEHGIIJGI=FHIJJJHHHFEFFFFFEECEDDBBDCCCC@CCCDD:
@ERR599021.1952 PHOSPHORE:151:D0EFNACXX:2:1101:9783:2503/1
```

Here we can look at and analyze the quality lines:
For example:

`CCCFFFADHHHHHHJJJJJHIJJJJJJJJJJJJJJJHHHFFEBDDDDD`
`CCCFFFFFHHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJIJJJJIIJEHG`

- Character `C` → ASCII 67 → 67-33=34 → quality = 34 (error probability ≈ 0.0004 = 0.04%).
- Character `F` → ASCII 70 → 70-33=37 → error probability ≈ 0.0002.
- Character `J` → ASCII 74 → 74-33=41 → error probability ≈ 0.00008 (very high quality).

In the example there are many `H`, `I`, `and J` →, which means high quality data.

```
(tp_tara) svetlana@DESKTOP-LAL2IVC:~/Metagenomique$ wc -l reads.fastq
400000 reads.fastq
```

The file `reads.fastq` contained 400,000 lines, corresponding to 100,000 retained sequences.

Q1. How many sequences were retained?

```
(tp_tara) svetlana@DESKTOP-LAL2IVC:~/Metagenomique$ head aligned.fq
@ERR599021.226 PHOSPHORE:151:D0EFNACXX:2:1101:7772:2071/1
GCACACGCACACGCACACGCACACGCACACGCACACACACACGCACACGCACACGCACACGCACACGCAGACGCAGACGCACACGCAGACGCAGA
+
CCCFFFADHHHHHJJJJJHIJJJJJJJJJJJJJJJHHHFFEBDDDDD=B?=B?@@8@BB57B758;5805<9@D&08>B>5<B9>B-9<<@>&5?9<B39
@ERR599021.930 PHOSPHORE:151:D0EFNACXX:2:1101:6892:2426/1
AGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAATTTCTTCAAGAGTTTGTCCCCAGAATTTCTCCCCAGAATTTCTCCACGG
+
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJIJJJJJJJJIJJJJJIIJJEHGIIJGI=FHIJJJHHHFEFFFFFEECEDDBBDCCCC@CCCDD:
@ERR599021.1952 PHOSPHORE:151:D0EFNACXX:2:1101:9783:2503/1
GGAAGGTCCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCGCGTAGGTGGTTTATTAAGTTGGATGTGAAATCCCCGGGCTCAACCTGGGAACTGCATC
(tp_tara) svetlana@DESKTOP-LAL2IVC:~/Metagenomique$ wc -l aligned.fq
(tp_tara) svetlana@DESKTOP-LAL2IVC:~/Metagenomique$ wc -l aligned.fq
392372 aligned.fq
```

The file `aligned.fastq` contained 392,372 lines, corresponding to 98,093 retained sequences (4 lines per FASTQ record) compared to 100,000 before alignment. Consequently, the 1,907 sequences were discarded during alignment.

We will now check the quality of the retained sequence data set. This is an essential step of an analysis like this one: indeed, if we use a data sett of poor quality, then the results we will get from them will not be reliable.

```
(tp_tara) svetlana@DESKTOP-LAL2IVC:~/Metagenomique$ fastqc aligned.fq
null
Started analysis of aligned.fq
Approx 5% complete for aligned.fq
Approx 10% complete for aligned.fq
Approx 15% complete for aligned.fq
Approx 20% complete for aligned.fq
Approx 25% complete for aligned.fq
Approx 30% complete for aligned.fq
Too many tiles (>2500) so giving up trying to do per-tile qualities since we're probably parsing the file wrongly
Approx 35% complete for aligned.fq
Approx 40% complete for aligned.fq
Approx 45% complete for aligned.fq
Approx 50% complete for aligned.fq
Approx 55% complete for aligned.fq
Approx 60% complete for aligned.fq
Approx 65% complete for aligned.fq
Approx 70% complete for aligned.fq
Approx 75% complete for aligned.fq
Approx 80% complete for aligned.fq
```

Q2. For which criteria have our sequences been given a "warn"? And a "fail"? Take a look at the plots provided for these categories. Do you think they warrant any concern, or can we continue with the analysis?
We observed two *warn* flags (*Sequence Length Distribution*, *Overrepresented Sequences*) and one *fail* (*Per Sequence GC Content*).

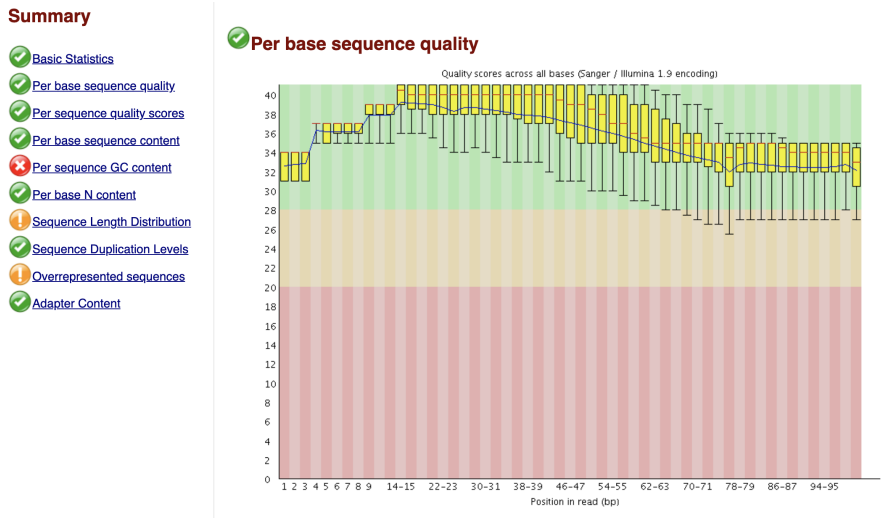


Figure 1: FastQC quality summary after alignment.

The 'GC content per sequence' criteria represents the distribution of the mean percentage of GC across all reads:

$$\%GC = \frac{G+C}{A+T+G+C} \times 100$$

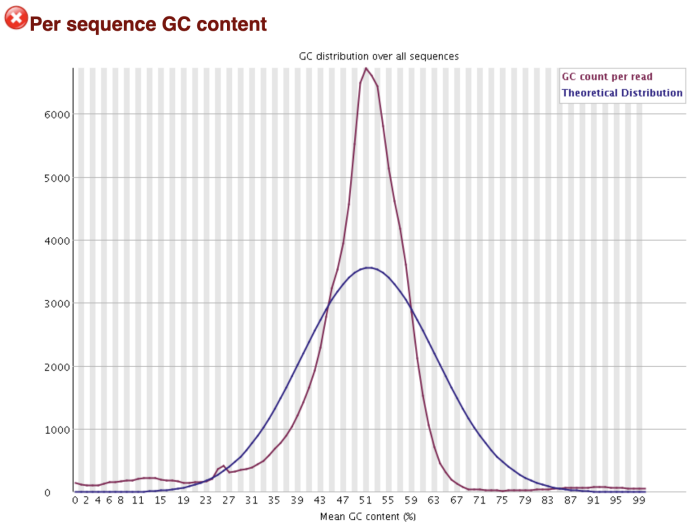


Figure 2: Comparison between the observed GC content of reads (red) and the theoretical normal distribution (blue).

The distribution of GC content deviated from the theoretical normal model (reduced variance). This indicates a slight bias in nucleotide composition, possibly due to amplification or sequencing

preferences. Although this does not invalidate the dataset, it suggests that some reads may be of uneven quality.

Concerning the Sequence Length Distribution, we notice that most of our sequences have a length of approximately 100 bp.

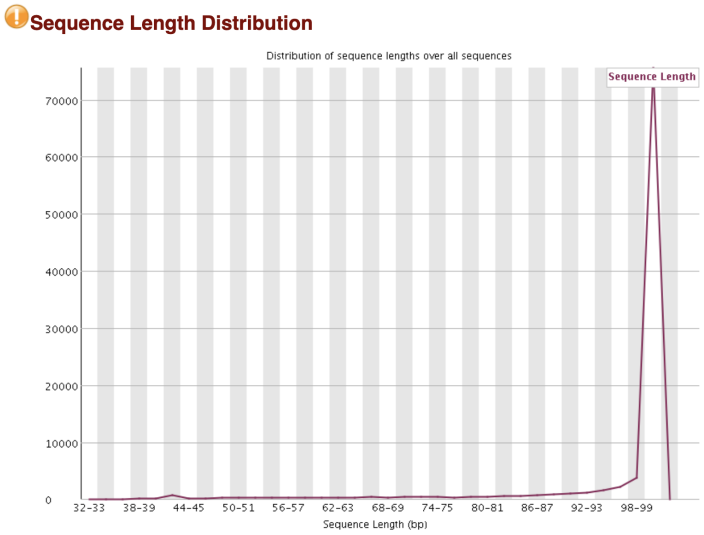


Figure 3: Distribution of read lengths over all sequences

Under ideal conditions, all fragments should have the same length. However, in our case, the variance remains small, indicating that the data set is consistent and suitable for further analysis.

For the Overrepresented Sequences criteria, several repetitive motifs (e.g., CACACA, GAGAGA not very complex motifs) appear at higher frequency, those are likely due to PCR amplification bias.

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CACACACACACACACACACACACACACACACACACACACACACACA	204	0.20943268381824529	No Hit
GAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGA	115	0.11806254234852064	No Hit
AGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG	107	0.1098494959242757	No Hit
TGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTG	106	0.1088228651212451	No Hit
ACACACACACACACACACACACACACACACACACACACACACACAC	102	0.10471634190912264	No Hit

Figure 4: Overrepresented sequence motifs

Overall, the data are of sufficient quality to continue, but these observations justify applying a filtering step to remove low-quality or biased reads and to ensure that only high-confidence sequences are kept for further analyzes.

Quality filtering

We will now filter the reads to eliminate those of poor quality. For this purpose, we will use the tool `Sickle`, in order to remove short sequences, those with an unsatisfactory quality score, and to potentially trim the beginning and/or end of sequences if this benefits their quality. We will apply the following quality thresholds:

- discard sequences of less than 30 nucleotides in length
- discard sequences with an average quality score of less than 30.

```
(tp_tara) svetlana@DESKTOP-LAL2IVC:~/Metagenomique$ sickle se -f aligned.fq -t sanger -o trimmed.fq -q 30 -l 30

FastQ records kept: 90779
FastQ records discarded: 7314

(tp_tara) svetlana@DESKTOP-LAL2IVC:~/Metagenomique$
```

Q3. How many sequences were removed by the filtering process?
After trimming with Sickle, 90,779 reads were retained, and 7,314 reads were discarded due to low quality or insufficient length.

Q4. Has there been a substantial change in sequence quality after trimming?
There were no significant changes in the overall sequence length or GC content distributions after trimming.

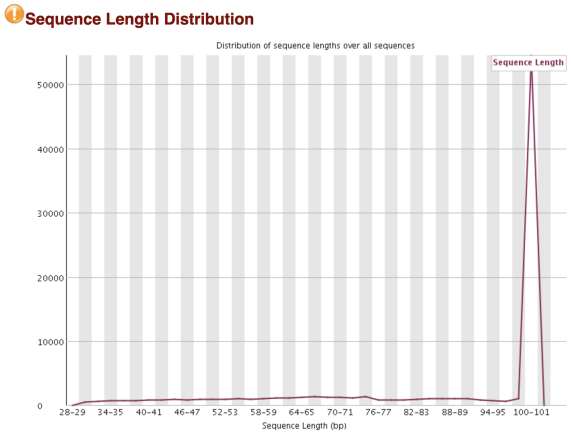


Figure 5: Sequence length distribution after trimming.

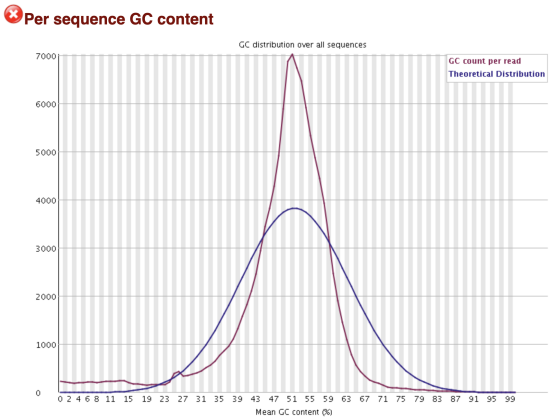


Figure 6: GC content distribution after trimming.

However, the number of overrepresented sequences decreased, with only the repetitive motif "CACACA..." remaining and at a lower frequency.

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CACACACACACACACACACACACACACACACACACACACACACACA	154	0.17077335935594046	No Hit

Figure 7: Overrepresented sequence motifs

It should be noted that there has been a minor change in the *Per sequence quality scores* plot. The distribution in aligned.fq was also good, but slightly more "stretched out". In the trimmed.fq, the peak is narrower, there is less variation, and most reads are concentrated in a narrow, high-quality range. After trimming, dataset looks even better and more stable than the previous one.

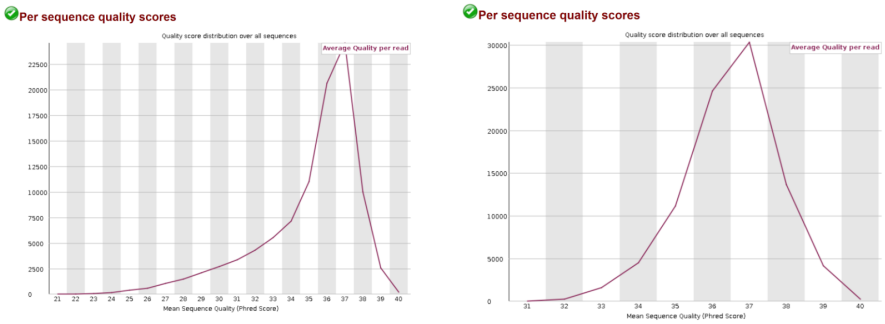


Figure 8: Quality score distribution before and after trimming.

A visible improvement was observed in the per base sequence quality plot, all positions in the reads are now in the green quality zone, indicating uniformly high-quality bases across the sequences. This confirms that the trimming step successfully removed low-quality and redundant reads, resulting in a cleaner and more reliable data set. The data are now of sufficient quality to proceed confidently with the next analysis steps.).

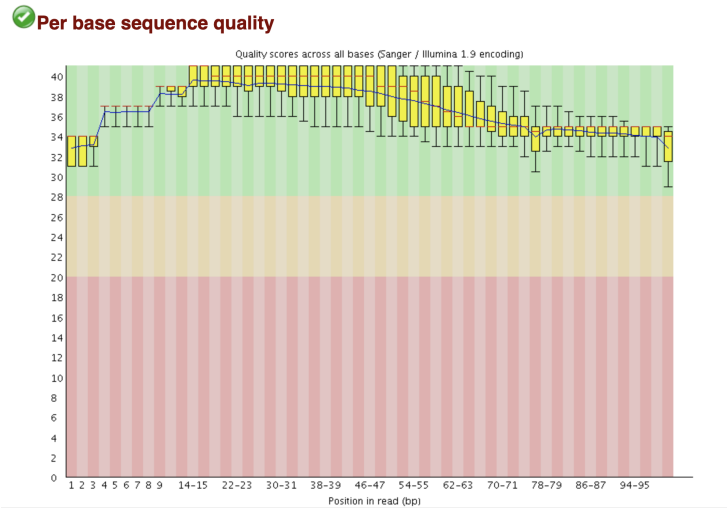


Figure 9: FastQC quality summary after trimming.

OTU clustering

Q5. OTUs defined and multi-sequence OTUs.

```
Reading file trimmed.unique.fa.sorted.fasta.temp 100%
7520774 nt in 86448 seqs, min 30, max 101, avg 87
Masking 100%
Counting k-mers 100%
Clustering 100%
Sorting clusters 100%
Writing clusters 100%
Clusters: 21716 Size min 1, max 1230, avg 4.0
Singletons: 16527, 19.1% of seqs, 76.1% of clusters
It took 13 seconds to cluster

Output File Names:
trimmed.agc.list
```

A total of 21,716 OTUs were identified after clustering. Among them, 16,527 OTUs were singletons (containing only one sequence), representing approximately 76.1% of all OTUs. This high proportion of singletons indicates a rich and complex microbial community, characterized by a large number of rare taxa. The remaining 5,189 OTUs contained multiple sequences, with the largest OTU comprising 1230 sequences.

Alpha diversity analysis

Once the OTUs have been defined, we can proceed to a number of biodiversity analyses, informing us on the biological diversity harbored in each of our samples (alpha-diversity) as well as the variation of this diversity across samples (beta-diversity).

Q6. Do higher index values indicate higher diversity?
For all three diversity indices (**Shannon**, **Inverse Simpson**, and **Chao1**), a higher value corresponds to a higher level of diversity.

- **Shannon index:** A higher Shannon value indicates that the environment is more unpredictable, with many species present and/or a more even distribution of individuals among species.
- **Inverse Simpson index:** A higher Inverse Simpson value means that individuals are more evenly distributed among species, resulting in a lower probability of randomly drawing two individuals belonging to the same species.
- **Chao1 index:** A higher Chao1 value reflects greater estimated species richness, accounting for both observed and unobserved (rare) species within the community.

```
mothur > summary.single(shared=trimmed.agc.shared, groupmode=t, calc=sobs-shannon-chao-invsimpson)

Processing group TARA_064_MES_IO

0.03

Processing group TARA_065_SRF_IO

0.03

Processing group TARA_152_MES_NAO

0.03

Processing group TARA_152_SRF_NAO

0.03

Output File Names:
trimmed.agc.groups.summary
```

Q7. Open the file trimmed.agc.groups.summary. Do the four indices of diversity agree on most and least diversified samples?

Table 1: Alpha-diversity indices per sample

Sample	Shannon	Chao1	Inverse Simpson
TARA_064_MES_IO	7.31	25 220	359.94
TARA_065_SRF_IO	7.38	28 258	297.82
TARA_152_MES_NAO	7.31	22 436	379.72
TARA_152_SRF_NAO	7.49	32 789	409.00

All four indices display a generally consistent trend: **TARA_152_SRF_NAO** (North Atlantic, surface) is the most diverse sample, whereas **TARA_152_MES_NAO** and **TARA_065_MES_IO** (mesopelagic samples from the North Atlantic and Indian Ocean) show the

lowest diversity. This pattern aligns with ecological expectations - **surface waters** typically harbor greater microbial richness and diversity than deeper layers, and overall, the **North Atlantic Ocean (NAO)** appears richer than the **Indian Ocean (IO)**.

Point out that Chao1 is sensitive to rare taxa and sampling effort, so a higher Chao1 (like in **TARA_152_SRF_NAO**) suggests more undetected rare species and greater potential richness.

However, the **Simpson index** introduces an important nuance. Unlike Shannon or Chao1, it accounts not only for species richness but also for the **evenness** of their distribution. According to this index, the mesopelagic samples (**MES_IO** and **MES_NAO**) appear more evenly structured - and thus more diverse - than **SRF_IO**. This suggests that while surface samples contain more species overall (as reflected by the Shannon and Chao1 indices), their communities are dominated by a few abundant taxa. In contrast, deep-sea communities display a more balanced distribution of individuals among species.

Let us now draw rarefaction curves for each of the samples, providing a visual representation of the number of detectable species in each sample according to the amount of sampled individuals.

```
mothur > rarefaction.single(shared=trimmed.agc.shared)

Using 12 processors.

Processing group TARA_064_MES_IO

0.03

Processing group TARA_065_SRF_IO

0.03

Processing group TARA_152_MES_NAO

0.03

Processing group TARA_152_SRF_NAO

0.03

It took 1 secs to run rarefaction.single.

Output File Names:
trimmed.agc.groups.rarefaction
```

Q8. How can you interpret this figure? Which samples are the most diversified? Consider the end slope of your curve: if we had sampled more individuals, do you think we would have found more OTUs or have we reached a maximum?

The rarefaction curves indicate that the surface samples (`SRF_IO` and `SRF_NAO`) are the most diverse, whereas the deep samples (`MES_IO` and `MES_NAO`) show lower species richness.

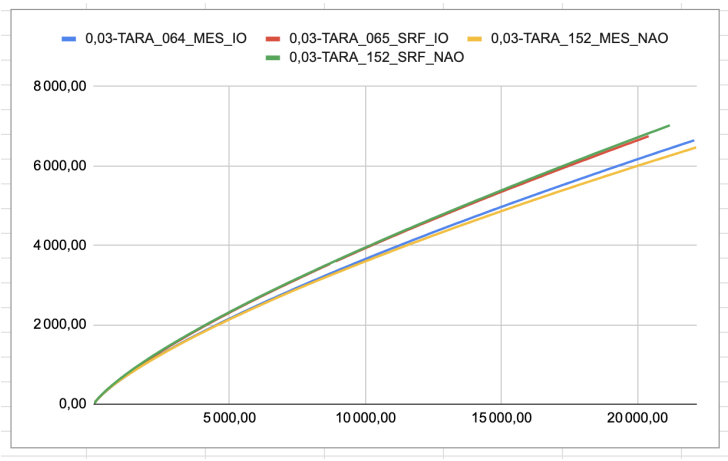


Figure 10: Rarefaction curves showing the number of observed OTUs (y-axis) as a function of the number of sequences sampled (x-axis).

The curves have not yet reached a plateau, suggesting that the number of OTUs continues to increase with sequencing depth. Therefore, additional sequencing would likely uncover other taxa, indicating that the full microbial diversity of these samples has not yet been captured.

Beta-diversity analysis

Another type of diversity analysis, called beta-diversity, consists in comparing together the OTUs present in different samples of our study. Thus, we consider each pair of samples to assess whether their species content is relatively similar or radically different.

Beta-diversity is a distinct notion from alpha-diversity: two samples containing a large number of species (high alpha-diversity) can be either very similar together if they contain roughly the same species (low beta-diversity) or very different if the species that compose them are only rarely found in the other sample (high beta-diversity).

```
mothur > summary.shared(shared=trimmed.agc.shared, calc=sharedsobs-sorclass-jabund)


Using 12 processors.
0.03

Output File Names:
trimmed.agc.summary
```

Q9. Which two samples are the most similar? The least similar? Between two samples from the same ocean but different depths, or two samples from the same depth but different oceans, which ones are the most different?

Table 2: Beta-diversity indices between pairs of samples at a 3% OTU clustering threshold. *sharedsobs* indicates the number of shared OTUs; *sorclass* corresponds to the Sørensen similarity index; *jabund* corresponds to the abundance-weighted Jaccard index.

Sample 1	Sample 2	Shared OTUs	Sørensen	Jaccard-abundance
TARA_064_MES_IO	TARA_065_SRF_IO	1111	0.834	0.448
TARA_064_MES_IO	TARA_152_MES_NAO	2035	0.690	0.295
TARA_065_SRF_IO	TARA_152_MES_NAO	1096	0.834	0.493
TARA_064_MES_IO	TARA_152_SRF_NAO	1065	0.844	0.478
TARA_065_SRF_IO	TARA_152_SRF_NAO	1611	0.766	0.397
TARA_152_MES_NAO	TARA_152_SRF_NAO	1004	0.851	0.494

The most similar samples are **TARA_152_MES_NAO** and **TARA_152_SRF_NAO**, with the highest Sørensen index (0.85), both from the North Atlantic, indicating they share a similar species pool despite being from different depths.

The least similar samples are **TARA_064_MES_IO** and **TARA_152_MES_NAO**, with the lowest Sørensen index (0.690) and Jaccard-abundance index (0.295), reflecting strong dissimilarity between mesopelagic communities from different oceans.

Samples from the same ocean but different depths are relatively similar, while samples from different oceans (even at the same depth) are the most different.

To better understand the similarity relationships between our different samples, we can proceed to a couple of visualisations.

```
mothur > venn(shared=trimmed.agc.shared)

0.03

Output File Names:
trimmed.agc.0.03.sharedsobs.TARA_064_MES_IO-TARA_065_SRF_IO-TARA_152_MES_NAO-TARA_152_SRF_NAO.svg
trimmed.agc.0.03.sharedsobs.TARA_064_MES_IO-TARA_065_SRF_IO-TARA_152_MES_NAO-TARA_152_SRF_NAO.sharedotus
```

Q10. How many OTUs are common to all four samples? In this diagram, do you see any particularly surprising or remarkable values, and if so, between which samples?

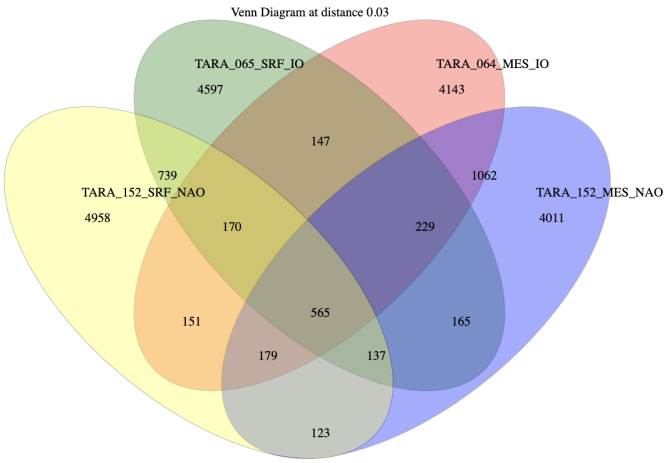


Figure 11: Venn diagram showing the number of shared and unique OTUs among the four samples.

Each sample also has a large number of unique OTUs, indicating high diversity and the presence of species specific to each location, reflecting the biogeographical and environmental variability of marine microbial communities.

A total of **565 OTUs** are common to all four samples, representing the core microbial community shared across both oceans and depths.

A particularly remarkable result is that the two mesopelagic samples, **TARA_064_MES_IO** and **TARA_152_MES_NAO**, share the largest number of OTUs - approximately **1856 OTUs** (1062 + 229 + 565). This is surprising because they come from different oceans, yet they show a strong overlap in species presence.

This may suggest that mesopelagic environments across oceans host similar microbial taxa, possibly due to comparable ecological conditions (low light, stable temperature, nutrient availability).

Q11. Which samples are grouped together? According to this tree, does biodiversity change more significantly based on the ocean of sampling or the depth of sampling?

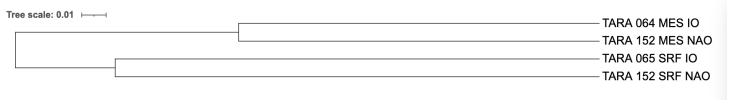


Figure 12: Hierarchical clustering of samples based on community composition

According to the clustering tree, samples group primarily by **depth** rather than by **ocean**. The two surface samples (`TARA_065_SRF_IO` and `TARA_152_SRF_NAO`) cluster together, as do the two mesopelagic samples (`TARA_064_MES_IO` and `TARA_152_MES_NAO`).

This pattern indicates that **depth has a stronger influence** on microbial community composition than the geographic origin of the sample.

Phylogeny

Having now defined the OTUs present in our sequence data file, the next logical step is to create a phylogenetic tree of the corresponding species.

```
(tp_tara) svetlana@DESKTOP-LAL2IVC:~/Metagenomique$ python3 helper.py otu-rep trimmed.agc.list trimmed.fa
Representative sequences written at oturep.fa
```

This will create a file called oturep.fa, containing one representative sequence per OTU.

```
>Otu09516|ERR599021.930
AGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAATTTCTTCAAGAGTTTGTCCCCAGAAT
TTCTCCCCAGAATTTCTCCACGG
>Otu20822|ERR599021.7441
GGAGTGAGGGAGAGAGGGAGGGATAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG
>Otu06272|ERR599021.37009
CATCGAAAATCATCGAAAATCATGGGCGCGCATGCATACATACATACATACATACATACATACATACATACATG
CATACATACATACATACATACAT
>Otu09632|ERR599021.49675
AAATTTATCAAATTTTCTTTTTTTTTTGATTTTTTTGCTTAAAATTTTCGGGTTTTTTTTTTTTT
>Otu03741|ERR599021.49754
AGTATGACTATAAGTTCATATTTTCACTGTGTCGGGTCTGTAGCTCAGTTGGTTAGAGCGCACCCCTGATAAGGGTGA
GGTCGGTGGTTCAAATCCACCCA
>Otu08848|ERR599021.57831
ACACACACACACACACACACACACACACACACACACACACACACCGATGACGAGGTGGCAACCT
>Otu17244|ERR599021.76674
TGGGTCATTAAATGGATATAAATTGAACAGTTTAGTCTTTATGGCGCATTAAACTGATATATATATATATATATATATAT
ATATATATATATACACTTCTG
>Otu10918|ERR599021.116606
ATTTGTCGAGATATTTTTGGCAACACACACACACACACACACACACACACACACACACACACA
>Otu09206|ERR599021.118774
AGATTAACTCACTGTCACCGCCATTGTAGCACGTGTGTAGCCCTGCGCGTAAGGGCCATGATGACTTGACGTCATCCC
CACCTTCCTCCGG
```

Figure 13: Representative sequence per OTU.

The next step will be to complie the BLAST database of SILVA sequence dataset.

```
(tp_tara) svetlana@DESKTOP-LAL2IVC:~/Metagenomique$ makeblastdb -in silva_ssu_db.fasta -dbtype nucl -out silva_blastdb


Building a new DB, current time: 10/07/2025 14:16:44
New DB name:   /home/svetlana/Metagenomique/silva_blastdb
New DB title:  silva_ssu_db.fasta
Sequence type: Nucleotide
Keep MBits: T
Maximum file size: 1000000000B
Adding sequences from FASTA; added 191162 sequences in 9.2961 seconds.
```

We can now align our OTU sequences against this database.

```
(tp_tara) svetlana@DESKTOP-LAL2IVC:~/Metagenomique$  blastn -query oturep.fa \-db silva_blastdb \-out oturep_silva_aln.blastn \-evalue 0.001 -perc_identity 97 \-ou
tfmt "6 qseqid sseqid evalue pident qcovs" \-max_target_seqs 1 -num_threads 4
```

Q12. How many of our sequences have been retrieved in the SILVA dataset by our search? What can we say of our oceanic sequences that have not been retrieved in the SILVA dataset?

```
(tp_tara) svetlana@DESKTOP-LAL2IVC:~/Metagenomique$  blastn -query oturep.fa \-db silva_blastdb \-out oturep_silva_aln.blastn \-evalue 0.001 -perc_identity 97 \-ou
tfmt "6 qseqid sseqid evalue pident qcovs" \-max_target_seqs 1 -num_threads 4
(tp_tara) svetlana@DESKTOP-LAL2IVC:~/Metagenomique$ wc -l oturep_silva_aln.blastn
6220 oturep_silva_aln.blastn
```

A total of **6,220 sequences** were successfully retrieved in the SILVA database, out of **21,376 OTUs** initially detected. This means that only about **29%** of our representative sequences had a significant match in SILVA, while the remaining **71%** could not be assigned.

Several factors can explain this relatively low retrieval rate:

- Some sequences may originate from poorly studied microbial lineages absent from SILVA.
- The representative sequences chosen for each OTU might not be the most informative or complete.
- Technical artefacts reads can reduce alignment success.
- The search was performed against a subset of SILVA, leading to potential data loss.

Among the retrieved sequences, only about **740** were assigned to known taxa, while the rest were classified as *uncultured* or *unidentified*, further illustrating the extent of unexplored microbial diversity in the ocean.

We will now create the taxonomic tree of all species retrieved in our samples.

Q13.Hover your mouse over the tree branches corresponding to the highest abundance peaks. What do they correspond to? What can you make of that?

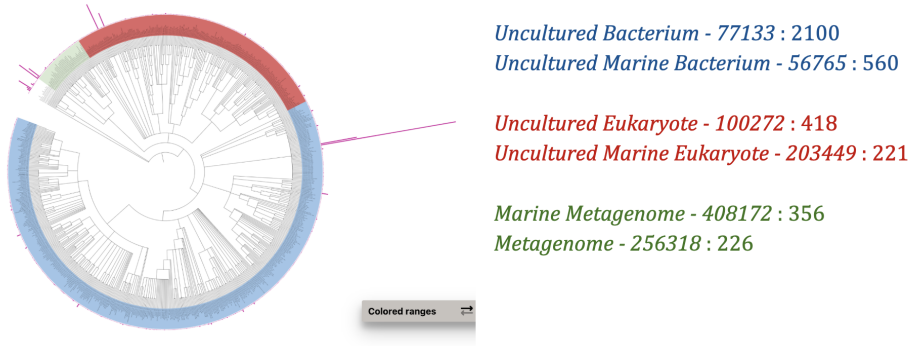


Figure 14: Phylogenetic tree showing the taxonomic affiliation and abundance peaks of dominant marine microorganisms.

The most abundant branches in the phylogenetic tree correspond mainly to `uncultured bacteria`, including `Uncultured Bacterium` and `Uncultured Marine Bacterium`, which dominate the dataset.

The predominance of “uncultured” taxa indicates that most of the organisms detected have not yet been successfully grown or characterized in laboratory conditions. This is particularly true for mesopelagic environments, which are difficult to study due to their extreme conditions (low light, high pressure, limited nutrients).

3 Short questions

A) What are the similarities and differences between the `FASTQ` and `FASTA` file formats?

Both `FASTA` and `FASTQ` are plain-text formats used to store nucleotide or protein sequences.

- **FASTQ** contains raw post-sequencing reads and includes a per-base Phred quality score. Each record has **four lines**:
  1. `@header` (sequence identifier)
  2. sequence
  3. `+` (separator)
  4. quality string (ASCII-encoded)
- **FASTA** contains only the sequence without quality information and is mainly used for reference genomes or processed sequences. Each record has **two lines**:
  1. >header
  2. sequence

Typically, raw reads (`FASTQ`) are aligned to a reference genome (`FASTA`) during bioinformatic analyses.

B) What is small subunit ribosomal RNA and why is it commonly used as an evolutionary marker?

The small subunit ribosomal RNA (**SSU rRNA**) is a structural and functional component of the small subunit of ribosomes in all living organisms: **16S rRNA** in prokaryotes and **18S rRNA** in eukaryotes. It plays a crucial role in mRNA recognition during protein synthesis.

It is widely used as an **evolutionary marker** because:

- It is present in all cellular life forms.
- It includes both conserved regions (for alignment across distant species) and variable regions (for distinguishing closely related taxa).
- It evolves slowly, allowing reconstruction of long-term evolutionary relationships.
- Its small size facilitates PCR amplification and sequencing.

C) What is an OTU, and what criterion is generally applied to group sequences together under a single OTU at the species level?

An **Operational Taxonomic Unit (OTU)** represents a cluster of DNA sequences grouped according to sequence similarity, typically approximating a taxonomic entity such as a species.

At the species level, sequences are commonly clustered into a single OTU if they share at least **97% nucleotide identity**.

D) Explain the difference between alpha- and beta-diversity.

- **Alpha-diversity** measures diversity *within* a single sample - it reflects both species richness (number of species) and evenness (distribution of abundances).
- **Beta-diversity** measures diversity *between* samples - it quantifies the degree of similarity or dissimilarity in species composition among communities.

E) What does a rarefaction curve represent and how is it interpreted?

A **rarefaction curve** shows the relationship between the number of sequences (or individuals) sampled and the number of species (or OTUs) observed. It helps assess whether the sampling effort was sufficient to capture the community's diversity.

Interpretation:

- At first, the curve rises steeply - each additional sequence reveals new species, meaning diversity is still being uncovered.
- If the curve continues to rise, more sequencing would be needed to capture the full diversity.
- If the curve reaches a plateau, sampling depth is considered sufficient to represent the community.

4 Interpretation of your results

Objectives :

The aim of this analysis was to explore the diversity and structure of marine microbial communities across different oceanic depths and regions using metagenomic data. Specifically, we aim to compare microbial diversity between surface (SRF) and mesopelagic (MES) samples collected from two distinct oceanic regions - the Indian Ocean (IO) and the North Atlantic Ocean (NAO). We further examined how environmental factors such as depth and geography shape community composition and similarity across samples, and identified the dominant taxonomic groups characterizing each environment.

Lower diversity in the mesopelagic zone compared to the surface

Rarefaction curves and alpha diversity indices (Shannon, Chao1, Simpson) show lower species richness and diversity in mesopelagic (MES) samples than in surface (SRF) samples.

Similar depth-related diversity gradients have been consistently reported in large-scale ocean microbiome surveys, which describe strong vertical structuring of microbial communities across light, temperature, and nutrient gradients (Sunagawa et al., 2015; Salazar et al., 2019).

This stratification reflects the contrasting environmental conditions between the surface and deeper layers of the ocean. The mesopelagic zone (200 - 1000 m depth) receives little or no sunlight, has colder temperatures, and limited nutrient availability, making it less favorable for photosynthetic organisms such as phytoplankton. Consequently, this layer supports fewer taxa and is dominated by slow-growing heterotrophic microorganisms adapted to low-energy conditions (Arístegui et al., 2009).

Depth stratification dominates community structure

Beta-diversity clustering (e.g., UPGMA tree) shows that samples from the same depth (MES or SRF) group together, regardless of their geographical origin (Indian Ocean vs North Atlantic).

This pattern highlights the strong influence of depth-related environmental gradients on microbial community structure. Factors such as light penetration, temperature, oxygen concentration, hydrostatic pressure, and nutrient availability are major ecological filters shaping microbial assemblages across the water column. These variables tend to vary more strongly with depth than with horizontal geographic distance, explaining why samples from the same depth but different oceans show greater similarity than samples from different depths within the same ocean.

This vertical stratification pattern has been consistently observed in global ocean surveys, where microbial composition and function display clear depth-dependent transitions rather than longitudinal ones (Zinger et al., 2011; Sul et al., 2013; Salazar et al., 2019). These studies support the idea that depth is the primary driver of microbial beta-diversity, structuring the distribution of taxa and their ecological roles in the marine environment.

Prevalence of uncultured lineages

A large proportion of OTUs detected in all samples are unclassified or uncultured taxa, even though some represent the most abundant lineages in the dataset.

This indicates that a significant part of marine microbial diversity remains poorly characterized. Many of these organisms likely inhabit extreme or specialized environments (e.g., deep or nutrient-poor mesopelagic waters), making them difficult or impossible to culture under standard laboratory conditions.

This observation aligns with global findings showing that the majority of microbial taxa in the ocean are still uncultured and undescribed, forming what has been termed the "microbial dark matter" (Amann Rosselló-Móra, 2016).

Recent advances in single-cell genomics (SCG) and metagenome-assembled genomes (MAGs) now allow access to genomic information from previously uncultured lineages, thereby overcoming the limitations of traditional cultivation methods. These approaches reveal the metabolic capabilities and ecological roles of microorganisms that cannot be grown under standard laboratory conditions (Tully et al., 2018; Pachiadaki et al., 2019).

5 Conclusions

Our analysis reveals clear ecological patterns structuring marine microbial communities. Surface waters (SRF) harbor richer and more diverse assemblages than deeper mesopelagic layers (MES), reflecting the strong vertical stratification of the ocean. Depth emerges as the main driver of microbial diversity, driven by gradients in light, temperature, oxygen, and nutrient availability.

These environmental factors define distinct ecological niches: photosynthetic and heterotrophic taxa dominate the dynamic, well-lit surface, while the mesopelagic zone supports slow-growing, metabolically specialized microorganisms adapted to low-energy conditions.

The predominance of uncultured lineages across all depths highlights how much of the ocean microbiome remains unexplored. Such uncharacterized diversity likely contributes significantly to global biogeochemical cycles, including carbon and nitrogen turnover.

Overall, our findings emphasize the key role of environmental gradients - particularly depth - in shaping microbial community structure and function. Metagenomic approaches remain essential to uncover the hidden complexity and ecological importance of marine microorganisms.

References

[1] Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., ... & Bork, P. (2015). Structure and function of the global ocean microbiome. *Science*, 348(6237). https://doi.org/10.1126/science.1261359

[2] Sul, W. J., Oliver, T. A., Ducklow, H. W., Amaral-Zettler, L. A., & Knight, R. (2013). Marine bacteria exhibit a bipolar distribution. *Proceedings of the National Academy of Sciences (PNAS)*, 110(6), 2342–2347. https://doi.org/10.1073/pnas.1212424110

[3] Zinger, L., Amaral-Zettler, L. A., Fuhrman, J. A., Horner-Devine, M. C., Huse, S. M., Welch, D. B. M., ... & Sogin, M. L. (2011). Global patterns of bacterial beta-diversity in the surface ocean. *Molecular Ecology*, 20(3), 529–541. https://doi.org/10.1111/j.1365-294X.2010.04988.x

[4] Amann, R. I., & Rosselló-Móra, R. (2016). After all, only millions? On the low estimated diversity of the marine biosphere. *mBio*, 7(4), e00999-16. https://doi.org/10.1128/mBio.00999-16

[5] Arístegui, J., Gasol, J. M., Duarte, C. M., & Herndl, G. J. (2009). Microbial oceanography of the dark ocean's pelagic realm. *Limnology and Oceanography*, 54(5), 1501–1529. https://doi.org/10.4319/lo.2009.54.5.1501

[6] Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H.-J., Cuenca, M., ... & Acinas, S. G. (2019). Gene expression changes and community turnover in the global ocean microbiome. *Cell*, 179(5), 1068–1083. https://doi.org/10.1016/j.cell.2019.10.014

[7] Tully, B. J., Graham, E. D., & Heidelberg, J. F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data*, 5, 170203. https://doi.org/10.1038/sdata.2017.203

[8] Pachiadaki, M. G., Brown, J. M., Brown, J., Bezuidt, O., Berube, P. M., Biller, S. J., ... & Stepanauskas, R. (2019). Charting the complexity of the marine microbiome through single-cell genomics. *Cell*, 179(7), 1623–1635. https://doi.org/10.1016/j.cell.2019.11.017