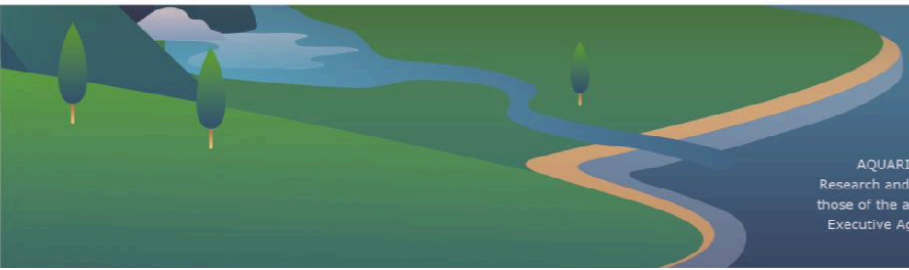
AQUARIUS initial Data Management Plan template forTACall applications

September 2024



Funded by the European Union

AQUARIUS has received funding from the European Union's Horizon Europe Framework Programme for Research and Innovation under grant agreement No 101130915. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

AQUARIUS initial Data Management Plan template for TA Call applications

As part of the Transnational Access (TA) projects in AQUARIUS, many new data sets in a large variety of data types will be collected by the TA scientific teams, making use and combining multiple and different observation installations as provided. There will be a strong effort in AQUARIUS to get the maximum return of investment from the TA activities towards serving the EU Mission and AQUARIUS Partnership targets and associated initiatives and projects with new data, data products, and scientific knowledge. Therefore, AQUARIUS has adopted an **open data policy**, which will be implemented with a **dedicated Data Management approach**, to ensure that all gathered and generated metadata and data will be managed in line with the FAIR principles (Findable, Accessible, Interoperable, Reusable). The metadata and data should become part of the archives managed and operated by leading European data management infrastructures for quality assurance, long term stewardship, and wide access and use.

TA scientific teams are advised to take note of the AQUARIUS Data Management approach document and guideline which can be found at:
https://zenodo.org/records/13642869

The AQUARIUS data management flow scheme includes a number of steps from planning to training to deployment to publishing, and a number of instruments, which should be applied in those steps by the TA scientific teams.

An initial step is that TA Call applicants are requested to complete an **initial Data Management Plan** for their proposed project, following the **AQUARIUS DMP template phase 1**, and to submit this as part of their TA Call application

Once the TA proposal has been awarded, the TA project team will be requested to review and refine the DMP once the TA team has a better understanding of the TA project that will be deployed and the AQUARIUS data management approach. AQUARIUS expert data centres can assist with the Call process and respond to questions about the AQUARIUS data management approach, for as far as answers cannot be found in the AQUARIUS Data Management approach document. In the phase after the TA project awarding there will be a kick-off meeting for selected TA projects with the AQUARIUS expert data centres to give support for completing and refining the TA project DMP template.

The questions in the DMP template are derived from the Horizon Europe DMP template and adapted for use as part of the AQUARIUS data management approach.

AQUARIUS initial Data Management Template - to be completed and submitted as part of the TA Call application

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The project aims to characterize microbial biodiversity and ecosystem functioning in Nordic marine environments, focusing on the Norwegian west coast. Metabarcoding (16S rRNA) and metagenomics samples, as well as physico-chemical and other biological parameters will be collected in order to identify the environmental and biological factors shaping microbial community composition and activity. These datasets directly support the project's objectives to establish spatial baselines of microbial diversity across Nordic seas, quantify microbial functional potential and bacterial respiration rates along the Norwegian coast, and integrate biodiversity monitoring with outreach through public aquaria. The resulting data will provide a comprehensive framework for assessing marine microbial diversity, ecosystem functioning, and responses to environmental change, forming the foundation for scientific publications and open-access dissemination.

What types and formats of data will the project generate/collect?

The project will collect and generate the following types and formats of data:

- **Physico-chemical characteristics of ocean surface waters**:
  - **Data**: Surface water temperature, salinity, oxygen, chlorophyll-a, colored dissolved organic matter (cDOM), turbidity, pH and partial pressure CO_2 (pCO_2) measured by the ferry box system onboard.
  - **Format**: Tabular ASCII CSV (provided by infrastructure operator).
- **Dissolved nutrients**
  - **Data**: Concentrations of dissolved nitrate, nitrite and phosphate.
  - **Format**: CSV.
- **Phytoplankton community data**
  - **Data**: Phytoplankton community composition based on morphology (10-200um)
  - **Format**: Zip-file with images (jpg) and corresponding metadata (JSON)
- **Bacterial traits (abundance, viability and autofluorescence)**
  - **Data**: Number of heterotrophic, autotrophic as well as viable cells per mL surface seawater.
  - **Format**: FCS (raw) and CSV (processed).
- **Bacterial respiration rate**
  - **Data**: Bacterial respiration rate in surface waters.
  - **Format**: CSV.
- **Genomic data - metabarcoding**
  - **Data**: Microbial community data (bacteria, archaea, eukaryotes and potentially fungi) through amplification of the 515Y/926R universal (and ITS region).

    - **Format**: FASTQ (raw reads), fasta (processed reads), TSV (taxonomic annotations and ASV count table).
- **Genomic data - metagenomic data**
    - **Data**: Genetic and metabolic potential and diversity of bacterial communities of selected indicative samples.
    - **Format**: FASTQ (raw metagenomes), DB including taxonomic and functional annotations (assembled genomes or metagenome-assembled genomes).
- **Metadata**
    - **Data**: Sampling date, time, location, depth, coordinates, devices, etc.
    - **Format**: JSON (following the MiXS water checklist[1]).

How is the original data gathered on the TA installation(s) and how do you plan to transfer it to you?

The original data for the project will be gathered using the ferry box and autosampler system onboard the vessel as well as through laboratory analyses of collected water samples. The ferry box will provide near-real time measurements of temperature, salinity, pH, oxygen, partial pressure of CO_2 (pCO_2), turbidity, chlorophyll-a and coloured dissolved organic matter (cDOM) every minute which relates to a data point approximately every 300-600 meters.

Water samples (approximately 4 L per station) for laboratory analyses will also be collected from either the ferry box directly or using the programmable autosampler system that is connected to the ferry box. We will gather samples to determine nutrient concentrations (nitrate, nitrite and phosphate), bacterial traits (cell counts, viability, autofluorescence and respiration rates), phytoplankton community composition as well to conduct eDNA analyses (metabarcoding and metagenomics). These samples will be preserved and transported to our home laboratory at Gothenburg University (GU) where they will be stored until further processing.

The sampling for bacterial traits, however, is dependent on the captain's approval to let us bring, store and utilize the necessary fixatives and chemicals. If the approval by the captain is not given, the project will proceed without bacterial trait and respiration rate analyses.

For flow cytometric (FC) determination of bacterial traits, 1800 mL water will be mixed with 200 mL paraformaldehyde (PFA) (10% paraformaldehyde + 0.5% glutaraldehyde) solution, incubated for 10 min in the dark, subsequently frozen and stored at -20°C until further processing. Per station, four replicates will be generated. Cell abundances and cell viability (and autofluorescence) will be determined using a Cube8 flow cytometer (Gothenburg University) or Cytoflex cytometer (Uppsala University).
For the assessment of bacterial respiration rates, we will conduct short, small-scale incubations using the BacLightTMRedoxSensorTM Green (RSG) Vitality Kit (Invitrogen) mainly following the manufacturers protocol. The bioassays will be performed at a few selected stations along the cruise track. For this, the 1 mL seawater (triplicate samples plus one negative control) will be incubated with 1 uL RSG close to its respective *in situ* temperature (within +/- 1.5 °C) for 30 min in the dark. Afterwards, samples will be fixed with

[1] https://www.ebi.ac.uk/ena/browser/view/ERC000024

100 uL PFA solution, incubated for 10 min in the dark and stored at -20 °C as described above. The measurements will be conducted using a Cytoflex cytometer (Uppsala University).

For eDNA analyses, we will filter 1 L (metabarcoding) or 2 L (metagenomics) of seawater through a 0.22 um Sterivex filter using a 60 mL serological syringe. For preservation, 1 mL of RNAlater-like solution will be added to the filter, before the filters are stored at -20 °C. DNA extraction and PCR amplification will be performed at our home laboratory at GU, while amplicon and shotgun sequencing will be conducted on Illumina platforms by SciLifeLab (Stockholm, Sweden).

For nutrient analyses, 15 mL seawater is taken up with a 20 mL serological syringe and filtered through a 0.2 um filter. The first 3 mL are used to rinse the filter and discarded, while the remaining 12 mL are filtered into a 15 mL falcon tube. The samples will be stored at -20 °C.

For phytoplankton community analyses, 100 mL seawater and 1 mL Lugol solution are mixed in a brown vessel and stored in the dark at around +5°C.

The raw data from the ferry box is available on the ferry box computer onboard. Automatic backups of the raw data to Google cloud are made daily and every 15 min for near-real time data as long as internet coverage is sufficient. Otherwise, automatic backups are postponed until next coverage. The automatic backups and data transfer from the TA installation for subsequent processing will be handled by NIVA, the operator of the research infrastructure. After data processing, NIVA will provide access for extraction of the raw and processed data to the PI and his team.

Results from laboratory analyses (nutrients and sequencing data) will be shared electronically by participating laboratories using secure data-sharing platforms.

What processing on the raw data do you plan? Please differentiate between data quality assurance (handling of outliers, missing and suspect values, null observations) and data harmonisation (code, label and QC flag explanations, consistent use of headers and units, data formatting, conforming to standards)

The ferry box data will be stored by NIVA using Google cloud. Upon ingestion, a couple of real time quality controls will be applied. These are defined in Copernicus Marine CMEMS (Jaccard et al., 2021; Von Schuckmann et al., 2010), though a few changes will be implemented to reduce the number of false positives.

Extra manual checks for outliers or suspect values are not covered by the described documentation. They can be provided by the research infrastructure operators on demand but the extra costs for this service are not covered by the AQUARIUS grant.

The ferry box system logs data every minute which corresponds to a data point approximately every 300-600m. Individual missing data points carry little weight for our study where sampling stations are dozens of kilometers apart. Missing data over longer distances cannot be regenerated at a later time point since the missing data point will most probably be unrelated to the last and next known measures due to geographic and atmospheric differences. Interpolation could be applied, however, the outcome will likely not result in realistic values

due to the aforementioned changes in environmental conditions. If ferry box data is missing for a longer period of data, this scenario and possible solutions for it will be discussed with the research infrastructure operator.

The metabarcoding data will be processed using the nf-core/ampliseq pipeline (Ewels et al., 2020; Straub et al., 2024). This pipeline applies a set of processing steps including quality control (FASTQ, Babraham Bioinformatics, 2012), read trimming (Cutadapt, Martin, 2011) and amplicon sequence variant inference and taxonomic classification (DADA2, Callahan et al., 2016) The processed read data (ASV table, taxonomy table) will then be imported into R for subsequent downstream analyses. Metagenome data will be processed using the anvi'o platform (Eren et al., 2021). This includes quality control (illumina-utils, Eren et al., 2013), metagenome assembly using i.e., metaSPAdes (Nurk et al., 2017) or MEGAHIT (Li et al., 2015), gene calling (pyrodigal-gv, Camargo et al., 2024), gene taxonomy (KrakenUniq, Breitwieser et al., 2018) and read recruitment (Bowtie2, Langmead and Salzberg, 2012). The downstream analysis will be conducted with assembled metagenomes or metagenome-assembled genomes. Processing of the raw sequences and metagenomes will be performed on resources provided by the PDC Center for High Performance Computing at the KTH Royal Institute of Technology (KTH) in Stockholm.

During the measurement of bacterial ecophysiology traits via flow cytometry, we will control for quality by using calibration beads, internal standards (where applicable) and check for a steady flow of the flow cytometer. Based on these controls, gates and quality flags are defined and applied as a common procedure by the research group (Dlugosch et al., 2023; Gasol and Morán, 2016) using FCS express as analysis software.

Phytoplankton diversity will be measured using a planktoscope following the protocol of Perruchon et al., 2025. Prior to the measurement, different calibration steps will be conducted to ensure proper performance of the instrument. The detected phytoplankton will be mapped against validated libraries (EcoTaxa) as well as on-sample basis using microscopy and expert knowledge within our research group.

The nutrients will be analyzed by the laboratory of Kristineberg Center which is accredited by SWEDAC to analyze seawater samples. Quality control flags will be defined and applied based on nutrient profiles and statistical analysis of replicates, e.g. through Bonferroni outlier test (Fox and Weisberg, 2011).

With respect to data harmonization, we will generate templates for sampling and methodological metadata to ensure consistent use of headers and SI units as well as consistent data formatting and metadata reporting according to the FAIR principle and our research group's data management.

When do you plan to perform these processing steps by means of a concise planning?

Water samples will be collected during the AQUARIUS access period along the Norwegian West coast. Laboratory processing, including analysis of environmental parameters, DNA extraction, library preparation, and sequencing of amplicon and metagenomic samples at SciLifeLab, is planned to take place within six to nine months after sampling. Subsequent bioinformatic processing, integration with environmental and biological datasets, and

statistical analyses will be performed during the following year. Data synthesis, interpretation, and preparation of publications and data submissions to public repositories (ENA and SBDI for sequencing data; EMODnet for physico-chemical and biological parameters) are expected to be completed within two years after sampling.

What is the expected size (Megabyte to Terabyte range) of the data?

To be collected datasets are in the Gigabyte range.

To whom might your data be useful, outside your project?

The primary users are expected to be researchers working in the field of microbial ecology. We will actively communicate the datasets and scientific findings to the wider community, encouraging their reuse.

2. Making data FAIR

2.1 Making data findable, including provisions for metadata

The **AQUARIUS Data Management approach** is aimed at inclusion of your collected/generated data in leading European aquatic data management infrastructures such as **SeaDataNet**[2] (physics, chemistry, geophysics, geology, and biology), **EurOBIS**[3] (marine biodiversity), **ELIXIR-ENA**[4] (biogenomics), **ICOS-Ocean**[5] (carbon), and **Copernicus INSTAC**[6] (Near- Real-Time data). Inclusion will be mediated through expert data centres which are partners in the AQUARIUS project. They will give training and education to the TA teams about planned data management practices as part of AQUARIUS WP5. In addition, as part of AQUARIUS WP6, they will give support to the TA teams for the adoption of the preferred metadata and data standards, and ingestion of the resulting metadata and data in repositories that are nodes of the indicated European data management infrastructures. This way, the metadata and data from the TA projects will become FAIR and also available for EMODnet, Copernicus Marine, DTO developments, EOSC, and global data exchanges such as GEOSS and the digital ocean ecosystem which is being developed in the framework of the UN-IOC-Ocean Decade programme. In the AQUARIUS Data Management approach, it is recommended to make use of the online **Data Submission service of EMODnet Data Ingestion**[7] ~~and/or **SeaDataNet SeaNoe service**~~[8] for transfer of your metadata and data packages to the AQUARIUS expert data centres.

Paragraph 5.4 of the **AQUARIUS Data Management approach** document lists the minimum set of metadata that EMODnet Ingestion requests you to enter in order to document your data packages submission. In addition, Annex 1 of the document gives a list of items that should be documented during the TA data acquisition and processing activities as these are later needed, when transforming data to the standard metadata and data formats as required.

[2] https://www.seadatanet.org
[3] https://www.eurobis.or
[4] https://www.ebi.ac.uk/ena/
[5] https://www.icos-cp.eu/observations/ocean/otc
[6] https://marine.copernicus.eu/about/producers/insitu-ta
[7] https://www.emodnet-ingestion.eu
[8] https://www.seadatanet.org/Software/SEANO

What naming conventions for your data files will you follow?

Data files will follow the convention:

YYYYMMDD_PROJECT_SENSOR/PARAMETER_STATION_NUMBER_vXX

where:

- YYYYMMDD is the sampling date, formatted as year (four digits), month (two digits), and day (two digits).
- PROJECT is UNHIDE.
- SENSOR/PARAMETER identifies the instrument (e.g., ferrybox) or measured parameter
- STATION_NUMBER is the sequential station number, beginning with 01
- vXX is the version of the data file

Will you in principle be able to follow the AQUARIUS Data Management approach for metadata? What possible challenges do you foresee?

We will be able to follow the AQUARIUS Data Management approach for metadata and will record all required metadata according to the above-mentioned templates and standards. No major challenges are foreseen. Delays in metadata uploading may occur due to sequencing turnaround times, but these will be managed by maintaining close coordination with SciLifeLab and the AQUARIUS data management team.

How will you document all this information before submission, especially lineage information (i.e. processing and QC steps)?

All sampling, laboratory, and analytical steps will be documented throughout following the AQUARIUS Data Management approach, specifically Annex 1, which provides a checklist for documenting data acquisition and processing during TA projects. This will include detailed logs for data acquisition methods, sensor specifications, calibration procedures as well as utilized softwares and tools. Data lineage and quality control steps for sequencing data will be recorded through version-controlled analysis pipelines (e.g., nf-core/Ampliseq, Anvi'o) and stored together with the corresponding data files. For each complementary environmental parameter, an individual file will document key information, including QC and data processing details. The documentation will be organized in a format compatible with ENA, SBDI and EMODnet standards, respectively, enabling a smooth submission process.

2.2. Making data openly accessible

The AQUARIUS project follows an **open data access policy.** Therefore, the **AQUARIUS Data Management approach** is aimed at inclusion and publishing of your collected/generated data as open metadata and data in trusted repositories, within the lifetime of the AQUARIUS project. However, temporary embargos (max 2 years after collection) are possible for specific data that require additional processing and to allow for scientific analyses that will lead to scientific papers.

Will you collect/generate data that might require a temporary embargo? Indicate which type of data and motivate its temporary embargo for each case.

Sample metadata will be submitted to BioSamples as soon as possible after the cruise. Genomic and physico-chemical data will be deposited with an embargo period of two years into public databases as soon as they are analysed.

Do you also plan to make the data and metadata available on another repository next to the AQUARIUS repositories, for instance an institutional, national or general data repository?

We plan to upload the sequencing data with contextual metadata (MiXS water checklist) to the European Nucleotide Archive (ENA). Samples will be directly registered at BioSamples and its accession numbers will be used for submission to ENA. Physico-chemical and biological parameters (bacterial trait data) will be uploaded either to EMODnet or SEANOE. Phytoplankton community data will be uploaded to EcoTaxa.

In case of use of exotic data formats for original data, what methods or software tools are needed to read and understand the data and could you provide documentation about that (including possible open-source code)?

Not applicable.

2.3. Making data interoperable

The **AQUARIUS Data Management approach** is aimed at making the collected/generated data interoperable by adopting common standards from leading European data management infrastructures. Use will be made of standard formats, QA-QC procedures, and controlled vocabularies for many metadata tags. For this process the TA teams will be given support by the AQUARIUS expert data centres, who also can help with use of common tools and services. To be able to do these transformations from original metadata and data to standardised metadata and data, it is important that relevant information about the data collection/generation is noted down during the actual events. This consists of the **AQUARIUS TA Data Summary Log** which should be maintained by the PI of the TA project scientific team to keep an overview and index of the data collection events.

It should be shared afterwards with the AQUARIUS data centre experts as it will provide an index to the collected/generated data and relevant documentation, such as who, when, where, how, etc. It contains only metadata and no data. For keeping this log, the PIs of the TA projects will receive a tablet from WP6 with a preloaded app, that will facilitate entering, browsing, and exporting information. The tablet should be returned to the WP6 team after finalisation of the TA acquisition activities. The **AQUARIUS TA Data Summary Log** will serve as a list for the expert data centres to know what data to expect from where and who and as a checklist for the following steps in the data management workflow.

In addition to the TA Data Summary Log it is recommended that the researchers of the TA teams document several aspects of the data acquisition and following processing, so that the later transformation to the common standards of the European marine data management infrastructureswillbeeasier. Annex 1 of the AQUARIUS Data Management approach

document gives a checklist which documentation TA researchers are recommended to maintain for this purpose.

How are you going to organise gathering the input required for the AQUARIUS TA Data Summary Log during the TA activities?

All information relevant to data collection and generation will be documented throughout the TA activity and the subsequent data processing phase in the AQUARIUS TA Data Summary Log file. The AQUARIUS App will be used for entering, browsing, and exporting the relevant information of the TA Data Summary Log. Metadata collection will follow the predefined structure and mandatory fields within the AQUARIUS app and Annex 1 of the AQUARIUS Data Management Approach. Collected datasets will be registered in the AQUARIUS App as close as possible to the time of acquisition using our research groups tablet. Both participants of the TA activity will be trained for maintaining the data log and entering information into the AQUARIUS app. The completeness and consistency of the metadata will be reviewed at the end of each sampling day.

Will you in principle be able to organise your TA team during the TA activities in order to gather the input required for the AQUARIUS TA Data Summary Log? What possible challenges do you foresee?

I will be able to organize my TA team (consisting of me and one colleague) during the TA activities in order to gather the input required for the AQUARIUS TA Data Summary Log. The only anticipated challenge concerns the use of a programmable autosampler, which will collect samples automatically during the night. As these samples will be processed the following morning, the corresponding entries in the Data Summary Log will also be completed at that time and not during the night when the samples are collected.

2.4. Increase data re-use (through clarifying licences)

The AQUARIUS project follows an **open data access policy,** and the AQUARIUS Data Management approach is aimed at inclusion and publishing of your collected/generated data as open metadata and data in trusted repositories, within the lifetime of the AQUARIUS project. The preferred licence for use of AQUARIUS data is CC-BY-4.0.

Do you foresee any difficulty for your collected/generated data to follow this policy?

No, I do not foresee any difficulties in ensuring that the data collected and generated through this project comply with the AQUARIUS open data access policy. All data and associated metadata will be published in trusted repositories under the CC-BY-4.0 license in accordance with the AQUARIUS Data Management approach.

3. Other research outputs

Do you plan to generate other research outputs next to the collected/generated data such as digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) results?

Our focus will be to publish our findings in open-access journals and to present them at national and international conferences and workshops. In addition, we plan to raise awareness

about the importance of microbial diversity and its role in marine ecosystems within the general public through our collaborative network with public aquaria as well as through local and regional outreach events by providing posters and giving presentations or student labs.

4. Allocation of resources

Terms & Conditions of AQUARIUS Funding require TA teams to follow the AQUARIUS Data Management Approach. This includes efforts for:

- updating their initial **Data Management Plan (DMP)** per TA project, as included in their original proposal submission, after being awarded and getting ready for the actual TA project deployment
- taking part in WP5 data management training activities to learn about AQUARIUS data management standards and getting experience with tools and services
- using the app during TA activities to complete the **AQUARIUS TA Data Summary Log**
- completing a **Cruise Summary Report (CSR)**[9] in case the TA project will involve scientific cruises with Research Vessels
- transforming the original metadata and data to the AQUARIUS prescribed standards, performing QA-QC, and making these FAIR and ready for inclusion in AQUARIUS repositories, thereby supported by AQUARIUS expert data centres.

Who will be responsible for data management in your TA project?

The PI of this project, Nicolai Verbücheln, will be responsible for coordinating the data management in the TA project. The research infrastructure (NorSOOP), which will collect the ferrybox data, will be responsible for providing this dataset in the correct format and metadata, following Aquarius recommendations and best practices.

5. Data security

The operators of the repositories as used for the AQUARIUS archival and long-term stewardship have security provisions in use such as for backups, secure storage, firewalls, etc.

What provisions are in place by the TA teams for data security (including backups, secure storage and transfer) in the initial phase, before the data is transferred to the AQUARIUS expert data centres?

Data security in the initial phase will be ensured through multiple backups and secure storage solutions. All data will be stored locally on institutional computers and backed up regularly on ~~the University of Gothenburg~~'s OneDrive as well as on an external hard drive.

6. Ethical aspects

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?

[9] https://www.seadatanet.org/Metadata/CSR-Cruises

No ethical or legal issues are expected to impact data sharing. No research permits are required for this project, as the regulation on the export and utilization of genetic resources from nature in Norway related to the Nagoya Protocol has not yet entered into force (personal communication with the Norwegian Environmental Agency, 7th October 2025). Until further notice, neither a Prior Informed Consent (PIC) nor Mutually Agreed Terms (MAT) permit is necessary.

7. Other issues

Will you follow any other national/ sectoral/ departmental procedures for data management, on top of the infrastructure and procedures that AQUARIUS will provide as part of its data management approach?

Yes, I will follow the data management procedures used by our research group (e.g. running sample identifiers and protocols for eDNA metadata sharing) and according to the guidelines of the University of Gothenburg, Sweden.

8. References

Babraham Bioinformatics, 2012. FastQC - A Quality Control tool for High Throughput Sequence Data [WWW Document]. URL https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed 4.11.25).

Breitwieser, F.P., Baker, D.N., Salzberg, S.L., 2018. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biology 19, 198. https://doi.org/10.1186/s13059-018-1568-0

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13, 581–583. https://doi.org/10.1038/nmeth.3869

Camargo, A.P., Roux, S., Schulz, F., Babinski, M., Xu, Y., Hu, B., Chain, P.S.G., Nayfach, S., Kyrpides, N.C., 2024. Identification of mobile genetic elements with geNomad. Nat Biotechnol 42, 1303–1312. https://doi.org/10.1038/s41587-023-01953-y

Dlugosch, L., Bunse, C., Bunk, B., Böttcher, L., Tran, D.Q., Dittmar, T., Hartmann, M., Heinrichs, M., Hintz, N.H., Milke, F., Mori, C., Niggemann, J., Spröer, C., Striebel, M., Simon, M., 2023. Naturally induced biphasic phytoplankton spring bloom reveals rapid and distinct substrate and bacterial community dynamics. FEMS Microbiol Ecol 99, fiad078. https://doi.org/10.1093/femsec/fiad078

Eren, A.M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S.E., Schechter, M.S., Fink, I., Pan, J.N., Yousef, M., Fogarty, E.C., Trigodet, F., Watson, A.R., Esen, Ö.C., Moore, R.M., Clayssen, Q., Lee, M.D., Kivenson, V., Graham, E.D., Merrill, B.D., Karkman, A., Blankenberg, D., Eppley, J.M., Sjödin, A., Scott, J.J., Vázquez-Campos, X., McKay, L.J., McDaniel, E.A., Stevens, S.L.R., Anderson, R.E., Fuessel, J., Fernandez-Guerra, A., Maignien, L., Delmont, T.O., Willis, A.D., 2021. Community-led, integrated, reproducible multi-omics with anvi'o. Nat Microbiol 6, 3–6. https://doi.org/10.1038/s41564-020-00834-3

Eren, A.M., Vineis, J.H., Morrison, H.G., Sogin, M.L., 2013. A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology. PLOS ONE 8, e66643. https://doi.org/10.1371/journal.pone.0066643

Ewels, P.A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M.U., Di Tommaso, P., Nahnsen, S., 2020. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol 38, 276–278. https://doi.org/10.1038/s41587-020-0439-x

Fox, J., Weisberg, S., 2011. Multivariate linear models in R. An R Companion to Applied Regression. Los Angeles: Thousand Oaks.

Gasol, J.M., Morán, X.A.G., 2016. Flow Cytometric Determination of Microbial Abundances and Its Use to Obtain Indices of Community Structure and Relative Activity, in: McGenity, T.J., Timmis, K.N., Nogales, B. (Eds.), Hydrocarbon and Lipid Microbiology Protocols: Single-Cell and Single-Molecule Methods. Springer, Berlin, Heidelberg, pp. 159–187. https://doi.org/10.1007/8623_2015_139

Jaccard, P., Hjemann, D.O., Ruohola, J., Ledang, A.B., Marty, S., Kristiansen, T., Kaitala, S., Mangin, A., Ruohola, J., Ledang, A.B., Marty, S., Kristiansen, T., Kaitala, S., Mangin, A., 2021. Quality Control of Biogeochemical Measurements within Copernicus in situ TAC.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Li, D., Liu, C.-M., Luo, R., Sadakane, K., Lam, T.-W., 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12. https://doi.org/10.14806/ej.17.1.200

Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P.A., 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res 27, 824–834. https://doi.org/10.1101/gr.213959.116

Perruchon, A., Zaccomer, H., Fabien, L., 2025. PlanktoScope protocol for plankton imaging. Straub, D., Tångrot, J., Peltzer, A., Lundin, D., emnilsson, bot, nf-core, Sateesh_Peri,

Bennett, A., Sundh, J., DiegoBrambilla, Peer, A., E, T., Manoharan, L., Garcia, M.U., Minot, S., Clayton, D., Gabernet, G., Malladi, V., Patel, H., PhilPalmer, Vaulot, D., Ewels, P., Davenport, C., Leo, D.D., Hörtenhuber, M., dariader, Menden, K., 2024.

nf-core/ampliseq: Ampliseq Version 2.12.0. https://doi.org/10.5281/zenodo.14167648 Von Schuckmann, K., Garau, B., Wehde, H., Gies, T., Durand, D., Reseghetti, F., Wehde, H.,

Gies, T., Durand, D., Reseghetti, F., 2010. Real Time Quality Control of temperature and salinity measurements within MyOcean and Copernicus in situ TAC. https://doi.org/10.13155/74317