WP4, D4.2, D17 WORKFLOW FOR PCR-AMPLICON, METAGENOMES, AND TOTAL RNA ANALYSES OF GLACIER SAMPLES

Project Number:	10107276
Project name:	Center for Glacial Biome Doctoral Network
Project Acronym:	ICEBIO
Call:	HORIZON-MSCA-2021-DN-01
Topic:	HORIZON-MSCA-2021-DN-01-01
Type of Action:	HORIZON-TMA-MSCA-DN
Service:	REA/A/01
Project Start Date:	1 October 2022
Project Duration:	48 months
Deliverable Title:	Workflow for PCR-amplicon, metagenomes, and total RNA analyses of glacier samples
Deliverable Number:	D4.2
Type:	Document, report
Due date (month):	20
Lead Beneficiary:	AU
Dissemination Level:	PU – Public
Work Package No:	WP4
Lead Author:	Lars Gerrie van Dijk (AU)
Author(s):	Alessandro Sergio Cuzzeri Julien Cergneux
Reviewed by:	Alexandre Magno Barbosa Anesio
Approved by:	Alexandre Magno Barbosa Anesio

Funded by the European Union

ICEBIO

PREFACE

This document is intended to provide a general guide for a standardized workflow of PCR-amplicon, metagenomic, and metatranscriptomic (/total RNA) analyses of glacier samples within the ICEBIO consortium. The aim is to promote consistency and comparability across the studies. Through the adoption of the standardized workflows outlined herein, we hope for a seamless integration of data and insights between doctoral candidates, ultimately contributing to an improved understanding of the glacier biome.

INTRODUCTION

Sequencing technologies have emerged as one of the most valuable tools for microbial ecologists to study microbial communities in diverse ecosystems. It has provided scientists a way to tap into the genetic and functional diversity of microbial communities, uncovering their roles, interactions, responses, and ecological impacts. Particularly in the cryosphere (i.e. the frozen parts of Earth), sequencing methodologies have been essential to uncover this environment as a biologically diverse biome that is dominated by microbial life - *for an extensive review of the microbial diversity in the glacier biome, we recommend Anesio et al. (2017)*.

Microorganisms in the cryosphere biome have been found to play pivotal ecological roles in the local ecosystem and biosphere. They are the predominant drivers of biogeochemical cycling processes in these systems and have been found to be involved in the fixation and sequestration of carbon and nitrogen species in biomass that may accumulate in these icy regions (Irvine-Fynn et al., 2021; Segawa et al., 2014; Telling et al., 2011; Telling et al., 2012). On glaciers and ice sheets, meltwater may export these nutrients to downstream ecosystems with possibly profound effects (Kellerman et al., 2020; Stevens et al., 2022). Other known ecological impacts include enhanced surface melt through the discoloration of snow and ice surfaces by algae (Cook et al., 2020; Halbach et al., 2022; Hoham & Remias, 2020; Millar et al., 2024; Tedstone et al., 2017), and the formation of cryoconite holes (Aoki et al., 2018; Fountain et al., 2008; Pittino et al., 2018). Knowledge of microbial diversity and active microbial processes is primarily restricted to the surfaces of glaciers and ice sheets due to the difficulties involved in sampling subsurface environments.

The endeavors made to understand the biology within the cryosphere have only recently emerged, and knowledge of the active microbial processes remains severely limited in this understudied biome. Sequencing technologies and approaches are expected to remain the primary tools for exploring microbial communities in cryospheric ecosystems. The resolution depth makes sequencing approaches an attractive tool for studying low abundance samples, such as englacial environments, snow, and aerosols (Boetius et al., 2015; Stibal et al., 2015). In these environments, microbial abundance may vary between 10^1 and 10^6 cells per milliliter sample. For sequencing approaches, this is, however, not a severe limitation as recent advancements have made it possible to employ sequencing technologies at a single-cell level (Chijiiwa et al., 2020; Wang et al., 2023).

Within the ICEBIO consortium, several projects will employ sequencing approaches to study the diversity and activity of whole microbial communities in cryospheric ecosystems. The most popular sequencing approaches for whole microbial community analysis include PCR-amplicon sequencing, metagenomics, and metatranscriptomics. Each approach will be further introduced in more detail in separate sections in this document. In short, PCR-amplicon sequencing is used to obtain both taxonomic and functional insights by targeting specific genes of interest using primers. Metagenomics extends the information gain by sequencing all genomes within a sample, offering a broader view of functional capabilities within microbial communities. Neither approaches, however, shed light on the active processes of the microbial community. This gap is filled by metatranscriptomics, which not only catalogs microbial diversity but also captures the expression of active genes.

The raw data generated by these approaches can be difficult to process and analyze by scientists who lack the required competencies. Moreover, it can require a considerable time investment to do so. For this reason, (streamlined) bioinformatics workflows and pipelines have been developed by skilled bioinformaticians to ease the process. In this document, we present some of these workflows to be used by those within the ICEBIO consortium and beyond. An overview of the workflows can be found in Figure 1. We hope that these workflows will be adopted by the ICEBIO doctoral candidates, which will ease data sharing and comparability, and improve collaboration and the understandability of data processing between projects.

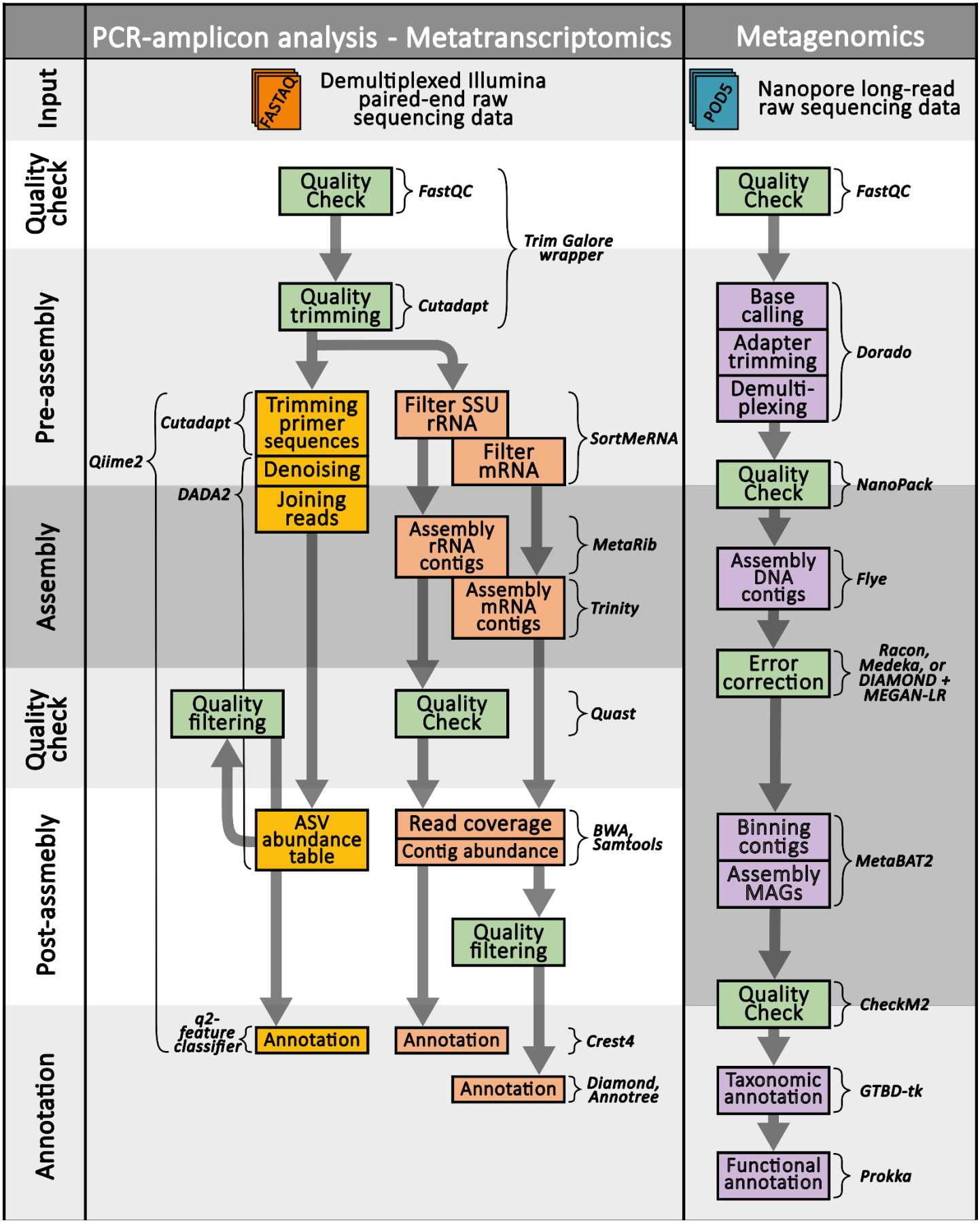

Figure 1. Graphical overview of bioinformatic workflows presented in this document for PCR-amplicon, metagenomic, and metatranscriptomic sequences. The annotated data output serves as the starting point for research objective-specific downstream analysis.

PCR-AMPLICON ANALYSIS

INTRODUCTION TO PCR-AMPLICON SEQUENCING

PCR-amplicon analysis offers a powerful method to investigate the presence of target genes in microbial communities, their abundance, and evolutionary relationships that can be used for taxonomic purposes. Popular targets for taxonomic diversity studies include the small subunit (SSU) 16S rRNA and the large subunit (LSU) 23S for archaea and bacteria, the SSU 18S rRNA and 28S for eukaryotes, and the internal transcribed spacer (ITS) regions for fungi (Schäffer et al., 2021). These genes are ideal targets for gaining insights into the diversity of the respective taxonomic groups of microbiome samples as these genes contain conservative and (hyper)variable regions that allow for targeting the gene and assessing the evolutionary relationships between genes, respectively.

In this section, we will focus on the 16S rRNA gene as our main target. The approximately 1550 base pairs long gene, contains nine hypervariable regions (V1-9), spanning from 30 to 100 base pairs each, which are interspersed with highly conserved regions (Figure 2). The highly conserved segments, act as anchors for universal primers that can be designed to target them for amplification of the targeted region by the Polymerase Chain Reaction (PCR). The primers can be barcoded with nucleotide tags, enabling multiplexing. This multiplexing strategy allows for simultaneous sequencing of multiple samples, thereby reducing costs and enhancing efficiency. The reason why only regions of the 16S rRNA gene are targeted and not the full length, is due to restrictions in the sequencing length up to a few hundred base pairs by the sequencing platform used. This restriction is an inherent issue of next-generation technologies that are predominantly used for this approach. However, full-length targeting and sequencing of the 16S rRNA gene and even the combined region of the 16S rRNA gene, ITSs region, and 23S RNA gene using third-sequencing technologies capable of long-read sequencing, are in active development.

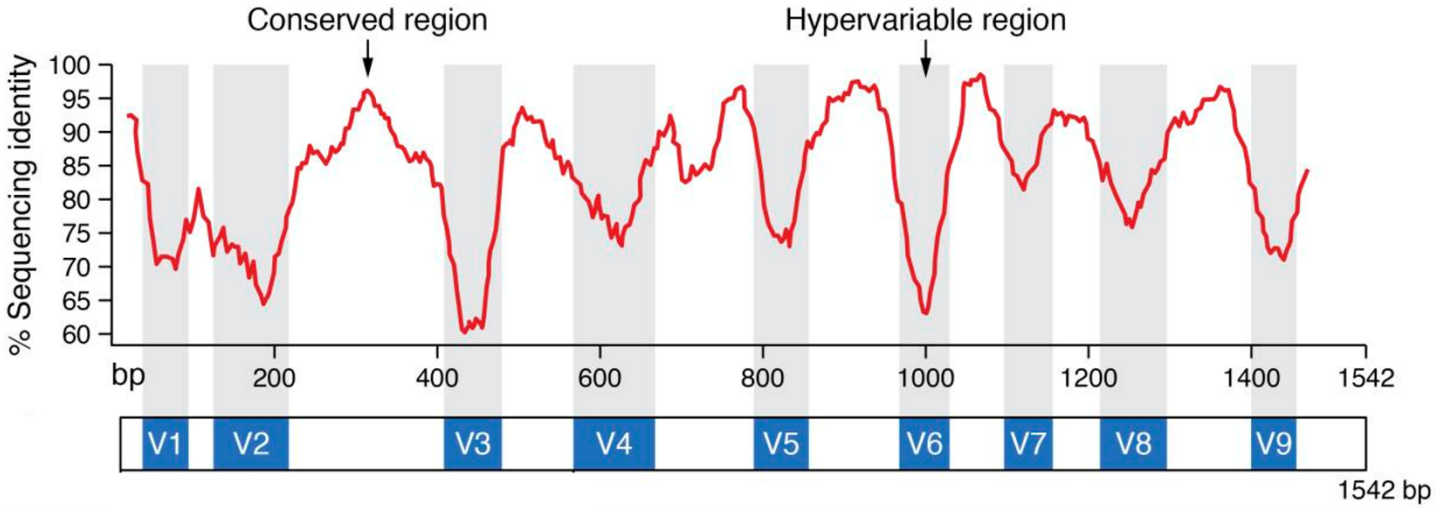

Figure 2. Percentage sequence identity of conserved and hypervariable regions of the bacterial 16S rRNA gene (adapted from Wensel et al., 2022).

When it comes to sequencing, next-generation Illumina platforms are the most widely used, providing high coverage and minimal per-base error rates for fragments under 300 base pairs, all at an economically low cost. The platform can be operated in two modes, single reads and paired-end reads. While the former offers cost-effective use, the latter sequences both ends of the fragments, not only increasing the number of reads but also enhancing read coverage, extending coverage of the target region, and significantly improving the accuracy of sequence alignment. This will eventually yield higher-quality amplicon sequence variants (ASVs), each a unique DNA sequence identified by even a single nucleotide difference, improving the analysis. Paired-end sequencing provides two file outputs per sample, one for the forward reads and one for the reverse. Before this, all pooled samples are demultiplexed by Illumina's software in one of the last steps in the sequencing workflow. The platform will therefore generate two FASTQ files per sample that contain the sequencing data. This data can be processed by the following bioinformatics workflow.

BIOINFORMATICS WORKFLOW

Quality filtering

Before taxonomic annotation, the raw sequencing reads are required to be processed through several steps, including filtering bad-quality reads and trimming bad-quality ends. QIIME2 is a free and open-source bioinformatics platform that is the main tool used in this workflow. Quality trimming is done before putting the data into QIIME2 as it has been shown that this improves the number of good-quality reads and abundance values (Mohsen et al., 2019). For quality trimming, we use the Perl wrapper Trim Galore (Felix Krueger, 2023) which combines the tools Cutadapt (Martin, 2011) and FastQC (Andrews, 2010) for quality trimming and generating HTML-based reports for inspection. Small reads (<200bp) are also removed.

QIIME2

Next, the data will be fed into the QIIME2 bioinformatics platform (Bolyen et al., 2019), which uses numerous plugins, including Cutadapt and DADA2 (Callahan et al., 2016) to process the quality-filtered reads to annotated ASVs. This is done in four stages:16S rRNA-specific trimming, denoising, joining of the reads into exact ASVs, and taxonomic annotation. The trimming step serves a different purpose than the quality trimming done before. This trimming step removes the primer sequences from the reads, which must be specified. After being further trimmed, the reads are denoised by the plugin DADA2, which identifies and separates true sequences from noise (i.e. sequencing errors and chimeras), after which the true sequencing reads are joined into ASVs. An additional step can be done to ensure the existence of only true ASVs in the data by comparing the ASVs across all samples and removing the ASVs that are not found in at least two samples.

Taxonomy assignment

Taxonomy is assigned to the ASVs using the feature-classifier plugin (Bokulich et al., 2018) in QIIME2 that trains a scikit-learn naïve Bayes classifier on a provided database. We use the Greengenes 99% OTUs and the SILVA SSU Ref NR99 databases for taxonomic assignment. However, other databases are available and may provide more useful results, albeit these two databases have presented reasonable results in microbiome analysis of glacial environments**.** As recommended by the developers, to improve the classification accuracy, the sequences in these databases are truncated to the region of interest, which is identified through an *in silico* analysis of the binding sites of the primers on the sequences in the databases.

Rarefaction curves

A final quality check is needed to determine if the sequencing depth was sufficient enough to capture the true diversity of a sample. This assessment is done by using the DADA2 alpha-rarefaction action that generates rarefaction curves. Rarefaction curves visualize the number of unique ASVs as a function of the number of sequences per sample, also called the sequencing depth. As the sequencing depth increases, the number of ASVs increases until a plateau is reached. A successful sequencing run will show a rarefaction curve reaching a plateau, meaning that the sequencing depth was high enough to capture (most of) the entire diversity of the sample. In other words, a higher sequencing depth would not increase the number of unique ASVs. On the other hand, if the curve does not reach a plateau, the sequencing run was unsuccessful in capturing the full diversity of the sample, and resequencing the sample with a higher sequencing depth may be required. For proper comparisons between samples, a similar sequencing depth is needed. Imbalanced sample sizes in the multiplexes DNA pool that is sequenced by the sequencing platform can result in a significant difference in the number of ASVs between samples. In this case, samples with a high amount of ASVs need to be rarefied to the minimum number of sequences found within samples. It is important to note that this process, although necessary for sample comparison, will result in the loss of a portion of the information. Alternatively, it is also possible to remove samples with too low sequencing depth to minimize this loss.

Downstream analyses

One of the data outputs is a (rarified) abundance table of the ASVs across the samples accompanied by the provided metadata. The analytical approach for this or any other data output from QIIME2 varies depending on the specific research questions and objectives of the researcher. Therefore, we will not provide a detailed description of this process in this document. A popular approach is to continue the downstream bioinformatic analyses in the

programming language of R. To achieve this, the package qiime2R can be used, which is able to convert QIIME2 objects into, for example, Phyloseq objects using the function `qza_to_phyloseq`. Phyloseq is an R package specifically developed for the analysis of microbial communities and proposes a set of tools for ecological and phylogenic analysis (McMurdie & Holmes, 2013).

METAGENOMIC ANALYSIS

INTRODUCTION TO METAGENOMICS

PCR-amplicon sequencing provides limited information on microbial communities but processing and analyzing its data remains relatively simple compared to metagenomic analysis. Sequencing all genomes of a biological sample generates vast amounts of data but can provide valuable insights into the genetic diversity, functional potential, and adaptability of microbial communities (Justice et al., 2008; Vandecraen et al., 2017). Metagenomic sequencing has become more widely used with the technological advancements that have revolutionized DNA sequencing, enabling more cost-effective and widespread approaches. With reduced costs, sequencing techniques have become accessible to a broader range of researchers and institutions (Christensen et al., 2015).

Widely used platforms for metagenomic sequencing include Illumina's second-generation sequencing platforms for massively paralleled short-read sequencing and Nanopore's third-generation sequencing technology for long-read sequencing of single DNA molecules. Illumina platforms have the advantage of providing low per-base error rates and relatively higher coverage, whereas Nanopore platforms can provide long sequencing reads of several kilobase pairs but at the expense of a higher error frequency for the assembly process. The drawback of short sequencing reads is the incapability of correctly handling repetitive and transposable elements in genomes for producing complete metagenome-assembled genomes (MAGs) that are of high quality. To overcome these limitations, a combination of both platforms can be used to obtain high-quality MAGs (Xia et al., 2023). However, third-generation sequencing technologies are improving at a rapid rate. Nanopore has been able to decrease its error frequency through new chemistries and flow cell designs and is leading a revolution to a new age of stand-alone long-read metagenomic analysis. Currently, high-quality near-complete metagenomes can be produced using the latest Nanopore flow cell R10.4.1 without "Illumina polishing" (Sereika et al., 2022).

Looking at these current developments, we advise the doctoral candidates within the ICEBIO consortium to solely employ long-read sequencing for metagenomic analysis using (preferably) Nanopore's latest technologies. For this reason, the bioinformatics workflow presented in this section for metagenomic analysis is focused on processing metagenomic sequencing data obtained by Nanopore platforms.

BIOINFORMATICS WORKFLOW

Compared to PCR-amplicon data, processing metagenomic data requires high processing capacity and memory storage. It is recommended that all steps in this workflow are run on a dedicated workstation or a High-Performance Computing cluster (HPC) through a job scheduler (e.g. SLURM). The workflow relies on open-source tools, the majority of them being available as Conda/Mamba installations, allowing for seamless handling of environments and the necessary dependencies. The workflow exists of four stages: initial processing of raw sequencing data (i.e. base calling, adapter trimming, and demultiplexing), quality control, assembly of MAGs, and finally, taxonomic and functional annotation of metagenomic content.

Base calling, adapter trimming, and demultiplexing

Nanopore sequencing platforms store their output in a POD5 file format (developed by Nanopore) that contains raw electric signal data from the sequencing platform. This data can be converted into written nucleotide sequencing reads in a process called base calling. The produced reads contain adapter and barcode sequences that were ligated at both ends of the reads prior to the sequencing and need to be removed post-sequencing. The adapter sequences are a requirement for the sequencing process, while the barcodes, similar to PCR-amplicon sequencing, are used for multiplexing samples to reduce costs and increase efficiency. After base calling, the adapter sequences are trimmed from the reads and demultiplexed into separate sample groups using the barcode sequences. The barcodes are removed during the process. In our workflow, all these steps are done using Nanopore's base caller

Dorado which has streamlined the process in a single command. However, if, for some reason, it is desired to do demultiplexing in a later step separately, the developers advise disabling adapter trimming as this can accidentally remove parts of the barcode sequences. To disable adapter trimming, the `--no-trim` flag should be included in the command. Trimming and demultiplexing can then be done with the command `dorado demux`. Specifying the `--emit-fastq` flag in this command will provide the standard output from Dorado, which are FASTQ files that are compatible with downstream analyses. Accidental removal of barcode sequences is prevented by Dorado when adapter trimming is done in combination with base calling and demultiplexing.

Quality control

To evaluate the quality of the reads in the FASTQ files, numerous packages and tools exist. For example, NanoPack, a Python-based package, contains several tools that have been specifically adapted to work with long-sequencing reads (De Coster & Rademakers, 2023). It includes the tools NanoPlot and NanoComp, which provide the user with graphical representations and statistical insights into the reads. A call to NanoPlot produces a list of plots and statistics, including the read length, quality, yield, N50 (the threshold of the minimum read size at which the longest reads cover 50% of the total length of all reads), and more. NanoComp creates, among other things, violin plots to compare the read lengths between different runs. Figure 3 shows two examples of the expected output.

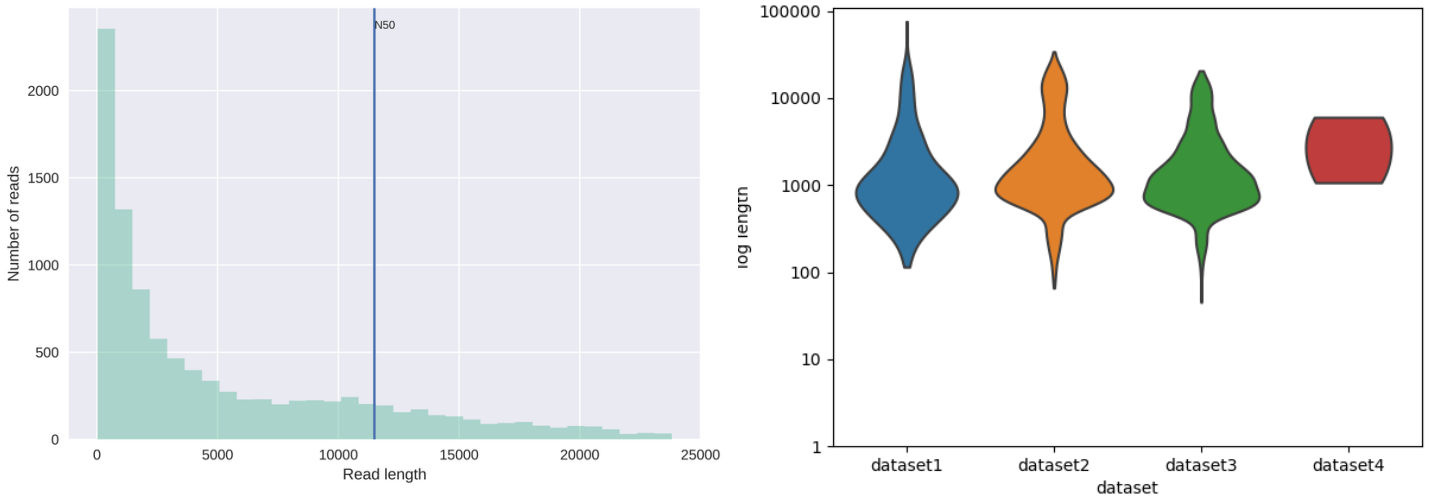

Figure 3. Example output from quality checks by NanoPlot (left) and NanoComp (right).

Assembly

This and the following step aim to assemble the (long) DNA fragments into the original DNA sequences of the organisms in the samples by merging overlapping reads first into contigs and then into MAGs. The assembly can be done for each sample separately or combined (co-assembly). The latter may improve the assembly process (not always) in samples that share taxa that differ in abundance and is recommended if the objective is to compare taxa abundances between samples.

To date, a number of different MAG assemblers are available for long reads (Latorre-Pérez et al., 2020). Our preferred choice is the tool Flye, which offers a good balance between quality and performance (Kolmogorov et al., 2019). The tool requires minimal parameter tuning and is straightforward to use (see documentation). Note that for metagenomic data, the flag `--meta` must be specified to enable the metagenomic mode. Other flags that may be important to include are `--plasmids` (to retrieve plasmids), `--threads` (number of parallel threads for processing), and `--iterations` (polishing step to correct small errors, default is one). However, like other long-read assemblers, Flye may struggle to obtain small plasmids (Johnson, Soehnlen, & Blankenship, 2023), and other tools, such as Unicycler, may be needed instead for this purpose. The output of Flye is an assembly file in FASTA format containing the assembled contigs, among other things.

Despite the polishing step in Flye to decrease the number of errors in the assembled contigs, it is advised to run a separate error correction step using different tools due to the high error profile of long reads obtained by third-

generation sequencers (Arumugam et al., 2023). We provide three ways of doing this. The simplest route is through the tools Racon (Vaser et al., 2017) and Medaka (developed by Nanopore). Both tools use the FASTA/FASTQ sequences containing the raw reads and the assembly output from Flye as input. The raw reads are in both tools re-aligned to the assembled contigs for assessing discrepancies using their own algorithms. Medaka also requires the raw signal output from the sequencer, which it utilizes to identify discrepancies in the sequences of the assembled contigs in a deep neural network model. The third option is possibly a more powerful approach but also more complex. It involves the concerted use of MEGAN-LR (Huson et al., 2018) and DIAMOND (Buchfink, Reuter, & Drost, 2021) to correct for frameshifts caused by indels (i.e. insertion-deletion mutations) in the sequences by producing DNA-to-protein alignments against the NCBI-NR database.

Binning and recovering MAGs (Metagenome-Assembled Genomes)
The final step is to group the corrected contigs into bins of which ideally each bin represents a single putative taxonomic group. A number of binning approaches exist that exploit similarities in sequence characteristics and sequencing data. In these approaches, the read coverage of contigs (i.e. the number of average reads aligned to a contig) often plays an important role as reads from the same species in a sample should have similar abundances, thus the read coverage for contigs made up from reads from the same species should, therefore, also be similar.

One such tool that exploits the read coverage is MetaBAT2, which we advise for binning in this workflow. This tool has been shown to outperform other binning tools in terms of performance and number of bins recovered at different completeness and precision cutoffs (Kang et al., 2019). It is, however, always encouraged to test different binning algorithms/programs to evaluate their performance, such as MaxBin2, CONCOCT and COCACOLA (Yue et al., 2020). MetaBAT2 requires the assembly file and a text file containing the read coverage of the contigs, which can be generated through the function `jgi_summarize_bam_contig_depths`. Alternatively, the tool minimap2 may be used, which is a mapping tool that is also used in the tool Racon. Minimap2 is a widely used mapping tool that performs well with long reads and large databases (Li, 2018). As input, it requires the assembled contigs and the raw read files. Note that the flag `--ax map-ont` has to be included. The output file is in a sequence alignment/map (SAM) file format that allows to store alignment coordinates for each read. MetaBAT2 requires the input file for the read coverage of the contigs in a binary alignment/map (BAM) format. The SAMtools package from Danecek et al. (2021) can be used to convert SAM files into BAM files. To convert the SAM files, include the sub-commands: `view` for conversion, `sort` for sorting, and `index` and `idxstats` for indexing and calculating the number of mapped reads per contig respectively. Besides submitting the input files, tuning the parameters in MetaBAT2 may be necessary to obtain high-quality MAGs. The developer's documentation provides several solutions for parameter tuning depending on the situation and we advise the user to consult the latest documentation for guidance on the best approach.

The output of MetaBAT2 is a FASTA file for each putative bin. It is worth knowing that MetaBAT2 is expected to produce better accuracy with increasing sample input due to the adaptive binning algorithm. As a final step, the quality of the bins is assessed using the well-known tool CheckM in its most recent release CheckM2 utilizes machine learning models to predict the quality of the recovered bins (Chklovski et al., 2023). As a "rule of thumb", high-quality MAGs are considered to be complete for 95% or more and containing less than 5% contamination.

Taxonomic and functional annotation
To assign the correct taxonomies to the MAGs, we make use of The Genome Taxonomy Database Toolkit (GTDB-tk) in this workflow (Chaumeil et al., 2019). The GTDB contains hundreds of thousands of genomes representing more than a hundred thousand unique bacterial and archaeal species. Species-level resolution is provided through average nucleotide identity, while higher taxonomic ranks are inferred from a set of more than a hundred marker genes using the HMMER algorithm (Eddy, 2011). The relevant command is `gtdbtk classify_wf`, which requires the input bins and the output paths. The reference database (~ 84 GB) must be downloaded beforehand.

Functional annotation of the MAGs is done by using the software tool Prokka (Seemann, 2014). Prokka requires the MAGs as input and produces a faa file containing the predicted protein sequences and a gff file containing both the MAGs and annotations among other things. The gff file can be directly loaded in genome visualization programs like

Artemis or, with slight modifications, in CIRCOS, to then proceed with visual MAG inspection. Depending on the research objectives, other downstream analyses may be done.

METATRANSCRIPTOMIC ANALYSIS

INTRODUCTION TO METATRANSCRIPTOMICS

PCR-amplicon sequencing and metagenomics have served as the predominant methodologies for whole microbial community analysis in natural environments in the past decades. Only recently has there been a substantial increase in metatranscriptomics projects, driven by the emergence of next-generation sequencing (NGS) technologies applied for analyzing metatranscriptomes (Shakya, Lo, & Chain, 2019). Metatranscriptomics succeeds its predecessors by providing insights into both the microbial diversity and the expression of active genes at any given time (Moran, 2009). Real-time insights into the gene expression profiles of microbial communities allow researchers to gain insights into the functional dynamics within ecosystems.

Within the cryosphere, metatranscriptomics has been applied to gain both taxonomic and functional insights into the active microbiome in thawing permafrost (Altshuler et al., 2019; Buelow et al., 2016; Coolen & Orsi, 2015; Scheel et al., 2023; Tveit et al., 2015; Varliero et al., 2021), glacial snow and ice (Bradley et al., 2023), perennial cave ice (Mondini et al., 2022), sea ice (Pearson et al., 2015), subglacial lakes (Gura & Rogers, 2020; Shtarkman et al., 2013), and cryoconite holes (Pittino et al., 2023; Segawa et al., 2020). In these studies, metatranscriptomics analysis has been used to elucidate microbial interactions (Scheel et al., 2023), the physiological state of microbial cells (Bradley et al., 2023), community responses (Coolen & Orsi, 2015), and biogeochemical cycling (Altshuler et al., 2019; Segawa et al., 2020).

The value of metatranscriptomics analyses in microbial ecology cannot be overstated but its rich and complex data is far from trivial to process and analyze, often demanding considerable time investment. To ease the data processing process, user-friendly streamlined bioinformatics pipelines have been developed, such as the TotalRNA pipeline from Campuzano (2023) that was developed at Aarhus University (Denmark). We encourage researchers within the ICEBIO consortium and beyond to adopt this pipeline to ensure uniformity of data processing.

BIOINFORMATICS WORKFLOW

The TotalRNA pipeline is continually undergoing development, and users are encouraged to refer to the latest documentation for guidance on utilizing the pipeline (Campuzano, 2023). Currently, the TotalRNA pipeline only handles raw paired-end reads from the Illumina platform. Here, a summary is given of the processing steps in the current version of the pipeline (v1.1.0). These steps are, however, automated within the pipeline, which is run by a single command and using a configuration file in which the user can specify instructions other than the default settings.

Summary of data processing steps

In the TotalRNA pipeline, first, raw paired-end Illumina reads are trimmed and quality-checked by Trim Galore (Felix Krueger, 2023). The trimmed reads are then categorized into four fractions based on their sequence similarity to reference databases: small subunit (SSU) rRNA reads, large subunit (LSU) rRNA reads, non-coding (nc) RNA reads, and unaligned reads (mRNA). The reference databases for sorting include the SILVA SSU Ref NR99 Release 138.1 (SSU rRNA), SILVA LSU Ref NR99 138 Release 138.1 (LSU rRNA) (Quast et al., 2012), and Rfam (Kalvari et al., 2020) (non-coding RNA) databases. While LSU rRNA and ncRNA reads are excluded from further processing, SSU rRNA and mRNA reads are assembled into RNA contigs by MetaRib (Xue, Lanzén, & Jonassen, 2020) and annotated. Crest4 (Lanzén et al., 2012) is used to annotate the reconstructed SSU rRNA contigs using a manually curated SILVA SSU database Release 138.1 (Bacteria, Archaea, and Fungi) and the PR2 v4.13 (Guillou et al., 2012) (Protist) database. Annotation of the mRNA contigs is done by DIAMOND (Buchfink, Xie, & Huson, 2015) using its internal database and the AnnoTree database (Gautam et al., 2022). Lastly, abundances of the RNA contigs are approximated by determining the coverage of the trimmed reads to the RNA contigs using BWA (Li & Durbin, 2009).

Data output

The pipeline provides several data outputs, including an annotated SSU rRNA and mRNA abundance table, a Phyloseq object of the SSU rRNA data that can be used for further analysis in the programming language of R, and an interactive Jupyter notebook in which the user can re-run parts of the code and change default parameters as desired. Many microbial ecologists are familiar with programming in R and the R package Phyloseq, therefore handling the Phyloseq object and the tabular data output for downstream analyses should be straightforward. All in all, the TotalRNA pipeline stands out as a robust solution that streamlines the complex data processing steps of metatranscriptomics data, helping researchers to more efficiently explore and interpret the taxonomic and functional dynamics within microbial communities.

CONCLUDING REMARKS

The goal of this document was to provide standardized bioinformatics workflows for PCR-amplicon, metagenomic, and metatranscriptomics analysis on glacier samples for the ICEBIO doctoral candidates and for anyone interested beyond. Through the adoption of these workflows, we hope to achieve increased comparability of data, improved understanding of data processing, and improved collaboration between the doctoral candidates of the ICEBIO consortium. These workflows, and especially the usage of the tools within, are by no means obligatory to follow. We urge anyone to always refer to the latest trends in these rapidly developing fields, and share and discuss these trends within the consortium.

REFERENCES

Altshuler, I., Ronholm, J., Layton, A., Onstott, T. C., W. Greer, C., & Whyte, L. G. (2019). Denitrifiers, nitrogen-fixing bacteria and N2O soil gas flux in high Arctic ice-wedge polygon cryosols. *FEMS Microbiology Ecology*, *95*(5). https://doi.org/10.1093/femsec/fiz049

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Anesio, A. M., Lutz, S., Chrismas, N. A. M., & Benning, L. G. (2017). The microbiome of glaciers and ice sheets. *npj Biofilms and Microbiomes*, *3*(1). https://doi.org/10.1038/s41522-017-0019-0

Aoki, T., Nagatsuka, N., Niwano, M., Sakaki, R., Shimada, R., Takeuchi, N., & Uetake, J. (2018). Temporal variations of cryoconite holes and cryoconite coverage on the ablation ice surface of Qaanaaq Glacier in northwest Greenland. *Annals of Glaciology*, *59*(77), 21-30. https://doi.org/10.1017/aog.2018.19

Arumugam, K., Bessarab, I., Haryono, M. A. S., & Williams, R. B. H. (2023). Recovery and Analysis of Long-Read Metagenome-Assembled Genomes. In S. Mitra (Ed.), *Metagenomic Data Analysis* (pp. 235-259). Springer US. https://doi.org/10.1007/978-1-0716-3072-3_12

Boetius, A., Anesio, A. M., Deming, J. W., Mikucki, J. A., & Rapp, J. Z. (2015). Microbial ecology of the cryosphere: sea ice and glacial habitats. *Nature Reviews Microbiology*, *13*(11), 677-690. https://doi.org/10.1038/nrmicro3522

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., & Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, *6*(1), 90. https://doi.org/10.1186/s40168-018-0470-z

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., . . . Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852-857. https://doi.org/10.1038/s41587-019-0209-9

Bradley, J. A., Trivedi, C. B., Winkel, M., Mourot, R., Lutz, S., Larose, C., Keuschnig, C., Doting, E., Halbach, L., Zervas, A., Anesio, A. M., & Benning, L. G. (2023). Active and dormant microorganisms on glacier surfaces. *Geobiology*, *21*(2), 244-261. https://doi.org/10.1111/gbi.12535

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(4), 366-368. https://doi.org/10.1038/s41592-021-01101-x

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59-60. https://doi.org/10.1038/nmeth.3176

Buelow, H. N., Winter, A. S., Van Horn, D. J., Barrett, J. E., Gooseff, M. N., Schwartz, E., & Takacs-Vesbach, C. D. (2016). Microbial Community Responses to Increased Water and Organic Matter in the Arid Soils of the McMurdo Dry Valleys, Antarctica [Original Research]. *Frontiers in Microbiology*, *7*. https://doi.org/10.3389/fmicb.2016.01040

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. *Nature Methods*, *13*(7), 581-583. https://doi.org/10.1038/nmeth.3869

Campuzano, C. (2023). AU-ENVS-Bioinformatics/TotalRNA-Snakemake: TotalRNA-Snakemake v1.1.0 (v1.1.0). *Zenodo*. https://doi.org/10.5281/zenodo.10068984

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, *36*(6), 1925-1927. https://doi.org/10.1093/bioinformatics/btz848

Chijiiwa, R., Hosokawa, M., Kogawa, M., Nishikawa, Y., Ide, K., Sakanashi, C., Takahashi, K., & Takeyama, H. (2020). Single-cell genomics of uncultured bacteria reveals dietary fiber responders in the mouse gut microbiota. *Microbiome*, *8*(1), 5. https://doi.org/10.1186/s40168-019-0779-2

Chklovski, A., Parks, D. H., Woodcroft, B. J., & Tyson, G. W. (2023). CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature Methods*, *20*(8), 1203-1212. https://doi.org/10.1038/s41592-023-01940-w

Christensen, K., Dukhovny, D., Siebert, U., & Green, R. (2015). Assessing the Costs and Cost-Effectiveness of Genomic Sequencing. *Journal of Personalized Medicine*, *5*(4), 470-486. https://doi.org/10.3390/jpm5040470

Cook, J. M., Tedstone, A. J., Williamson, C., McCutcheon, J., Hodson, A. J., Dayal, A., Skiles, M., Hofer, S., Bryant, R., McAree, O., McGonigle, A., Ryan, J., Anesio, A. M., Irvine-Fynn, T. D. L., Hubbard, A., Hanna, E., Flanner, M., Mayanna, S., Benning, L. G., . . . Tranter, M. (2020). Glacier algae accelerate melt rates on the south-western Greenland Ice Sheet. *The Cryosphere*, *14*(1), 309-330. https://doi.org/10.5194/tc-14-309-2020

Coolen, M. J. L., & Orsi, W. D. (2015). The transcriptional response of microbial communities in thawing Alaskan permafrost soils [Original Research]. *Frontiers in Microbiology*, *6*. https://doi.org/10.3389/fmicb.2015.00197

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

De Coster, W., & Rademakers, R. (2023). NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics*, *39*(5). https://doi.org/10.1093/bioinformatics/btad311

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Felix Krueger, F. J., Phil Ewels, Ebrahim Afyounian, Michael Weinstein, Benjamin Schuster-Boeckler, Gert Hulselmans, & sclamons. (2023). FelixKrueger/TrimGalore: v0.6.10 - add default decompression path (0.6.10). *Zenodo*. https://doi.org/10.5281/zenodo.7598955

Fountain, A. G., Nylen, T. H., Tranter, M., & Bagshaw, E. (2008). Temporal variations in physical and chemical features of cryoconite holes on Canada Glacier, McMurdo Dry Valleys, Antarctica. *Journal of Geophysical Research: Biogeosciences*, *113*(G1). https://doi.org/https://doi.org/10.1029/2007JG000430

Gautam, A., Felderhoff, H., Bağci, C., & Huson, D. H. (2022). Using AnnoTree to Get More Assignments, Faster, in DIAMOND+MEGAN Microbiome Analysis. *mSystems*, *7*(1), e01408-01421. https://doi.org/doi:10.1128/msystems.01408-21

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H. C. F., Lara, E., Le Bescot, N., Logares, R., . . . Christen, R. (2012). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(D1), D597-D604. https://doi.org/10.1093/nar/gks1160

Gura, C., & Rogers, S. O. (2020). Metatranscriptomic and Metagenomic Analysis of Biological Diversity in Subglacial Lake Vostok (Antarctica). *Biology*, *9*(3), 55. https://www.mdpi.com/2079-7737/9/3/55

Halbach, L., Chevrollier, L.-A., Doting, E. L., Cook, J. M., Jensen, M. B., Benning, L. G., Bradley, J. A., Hansen, M., Lund-Hansen, L. C., Markager, S., Sorrell, B. K., Tranter, M., Trivedi, C. B., Winkel, M., & Anesio, A. M. (2022). Pigment signatures of algal communities and their implications for glacier surface darkening. *Scientific Reports*, *12*(1), 17643. https://doi.org/10.1038/s41598-022-22271-4

Hoham, R. W., & Remias, D. (2020). Snow and Glacial Algae: A Review1. *Journal of Phycology*, *56*(2), 264-282. https://doi.org/https://doi.org/10.1111/jpy.12952

Huson, D. H., Albrecht, B., Bağcı, C., Bessarab, I., Górska, A., Jolic, D., & Williams, R. B. H. (2018). MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of

metagenomic long reads and contigs. *Biology Direct*, *13*(1), 6. https://doi.org/10.1186/s13062-018-0208-7

Irvine-Fynn, T. D. L., Edwards, A., Stevens, I. T., Mitchell, A. C., Bunting, P., Box, J. E., Cameron, K. A., Cook, J. M., Naegeli, K., Rassner, S. M. E., Ryan, J. C., Stibal, M., Williamson, C. J., & Hubbard, A. (2021). Storage and export of microbial biomass across the western Greenland Ice Sheet. *Nature Communications*, *12*(1), 3960. https://doi.org/10.1038/s41467-021-24040-9

Johnson, J., Soehnlen, M., & Blankenship, H. M. (2023). Long read genome assemblers struggle with small plasmids. *Microb Genom*, *9*(5). https://doi.org/10.1099/mgen.0.001024

Justice, S. S., Hunstad, D. A., Cegelski, L., & Hultgren, S. J. (2008). Morphological plasticity as a bacterial survival strategy. *Nature Reviews Microbiology*, *6*(2), 162-168. https://doi.org/10.1038/nrmicro1820

Kalvari, I., Nawrocki, E. P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., Rivas, E., Eddy, S. R., Finn, Robert D., Bateman, A., & Petrov, A. I. (2020). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, *49*(D1), D192-D200. https://doi.org/10.1093/nar/gkaa1047

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, *7*, e7359. https://doi.org/10.7717/peerj.7359

Kellerman, A. M., Hawkings, J. R., Wadham, J. L., Kohler, T. J., Stibal, M., Grater, E., Marshall, M., Hatton, J. E., Beaton, A., & Spencer, R. G. M. (2020). Glacier Outflow Dissolved Organic Matter as a Window Into Seasonally Changing Carbon Sources: Leverett Glacier, Greenland. *Journal of Geophysical Research: Biogeosciences*, *125*(4), e2019JG005161. https://doi.org/https://doi.org/10.1029/2019JG005161

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540-546. https://doi.org/10.1038/s41587-019-0072-8

Lanzén, A., Jørgensen, S. L., Huson, D. H., Gorfer, M., Grindhaug, S. H., Jonassen, I., Øvreås, L., & Urich, T. (2012). CREST--classification resources for environmental sequence tags. *PLoS One*, *7*(11), e49334. https://doi.org/10.1371/journal.pone.0049334

Latorre-Pérez, A., Villalba-Bermell, P., Pascual, J., & Vilanova, C. (2020). Assembly methods for nanopore-based metagenomic sequencing: a comparative study. *Scientific Reports*, *10*(1), 13588. https://doi.org/10.1038/s41598-020-70491-3

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094-3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads [next generation sequencing; small RNA; microRNA; adapter removal]. *2011*, *17*(1), 3. https://doi.org/10.14806/ej.17.1.200

McMurdie, P. J., & Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, *8*(4), e61217. https://doi.org/10.1371/journal.pone.0061217

Millar, J. L., Broadwell, E. L. M., Lewis, M., Bowles, A. M. C., Tedstone, A. J., & Williamson, C. J. (2024). Alpine glacier algal bloom during a record melt year. *Front Microbiol*, *15*, 1356376. https://doi.org/10.3389/fmicb.2024.1356376

Mohsen, A., Park, J., Chen, Y.-A., Kawashima, H., & Mizuguchi, K. (2019). Impact of quality trimming on the efficiency of reads joining and diversity analysis of Illumina paired-end reads in the context of QIIME1 and QIIME2 microbiome analysis frameworks. *BMC Bioinformatics*, *20*(1), 581. https://doi.org/10.1186/s12859-019-3187-5

Mondini, A., Anwar, M. Z., Ellegaard-Jensen, L., Lavin, P., Jacobsen, C. S., & Purcarea, C. (2022). Heat Shock Response of the Active Microbiome From Perennial Cave Ice [Original Research]. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.809076

Moran, M. A. (2009). Metatranscriptomics: Eavesdropping on Complex Microbial Communities. *Microbe*, *4*, 329-335. https://doi.org/10.1128/microbe.4.329.1

Pearson, G. A., Lago-Leston, A., Cánovas, F., Cox, C. J., Verret, F., Lasternas, S., Duarte, C. M., Agusti, S., & Serrão, E. A. (2015). Metatranscriptomes reveal functional variation in diatom communities from the Antarctic Peninsula. *The ISME Journal*, *9*(10), 2275-2289. https://doi.org/10.1038/ismej.2015.40

Pittino, F., Maglio, M., Gandolfi, I., Azzoni, R. S., Diolaiuti, G., Ambrosini, R., & Franzetti, A. (2018). Bacterial communities of cryoconite holes of a temperate alpine glacier show both seasonal trends and year-to-year variability. *Annals of Glaciology*, *59*(77), 1-9. https://doi.org/10.1017/aog.2018.16

Pittino, F., Zawierucha, K., Poniecka, E., Buda, J., Rosatelli, A., Zordan, S., Azzoni, R. S., Diolaiuti, G., Ambrosini, R., & Franzetti, A. (2023). Functional and Taxonomic Diversity of Anaerobes in Supraglacial Microbial Communities. *Microbiology Spectrum*, *11*(2), e01004-01022. https://doi.org/doi:10.1128/spectrum.01004-22

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590-D596. https://doi.org/10.1093/nar/gks1219

Schäffer, A. A., McVeigh, R., Robbertse, B., Schoch, C. L., Johnston, A., Underwood, B. A., Karsch-Mizrachi, I., & Nawrocki, E. P. (2021). Ribovore: ribosomal RNA sequence analysis for GenBank submissions and database curation. *BMC Bioinformatics*, *22*(1), 400. https://doi.org/10.1186/s12859-021-04316-z

Scheel, M., Zervas, A., Rijkers, R., Tveit, A. T., Ekelund, F., Campuzano Jiménez, F., Christensen, T. R., & Jacobsen, C. S. (2023). Abrupt permafrost thaw triggers activity of copiotrophs and microbiome predators. *FEMS Microbiology Ecology*, *99*(11). https://doi.org/10.1093/femsec/fiad123

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068-2069. https://doi.org/10.1093/bioinformatics/btu153

Segawa, T., Ishii, S., Ohte, N., Akiyoshi, A., Yamada, A., Maruyama, F., Li, Z., Hongoh, Y., & Takeuchi, N. (2014). The nitrogen cycle in cryoconites: naturally occurring nitrification-denitrification granules on a glacier. *Environmental Microbiology*, *16*(10), 3250-3262. https://doi.org/https://doi.org/10.1111/1462-2920.12543

Segawa, T., Takeuchi, N., Mori, H., Rathnayake, R. M. L. D., Li, Z., Akiyoshi, A., Satoh, H., & Ishii, S. (2020). Redox stratification within cryoconite granules influences the nitrogen cycle on glaciers. *FEMS Microbiology Ecology*, *96*(11). https://doi.org/10.1093/femsec/fiaa199

Sereika, M., Kirkegaard, R. H., Karst, S. M., Michaelsen, T. Y., Sørensen, E. A., Wollenberg, R. D., & Albertsen, M. (2022). Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat Methods*, *19*(7), 823-826. https://doi.org/10.1038/s41592-022-01539-7

Shakya, M., Lo, C. C., & Chain, P. S. G. (2019). Advances and Challenges in Metatranscriptomic Analysis. *Front Genet*, *10*, 904. https://doi.org/10.3389/fgene.2019.00904

Shtarkman, Y. M., Koçer, Z. A., Edgar, R., Veerapaneni, R. S., D'Elia, T., Morris, P. F., & Rogers, S. O. (2013). Subglacial Lake Vostok (Antarctica) Accretion Ice Contains a Diverse Set of Sequences from Aquatic, Marine and Sediment-Inhabiting Bacteria and Eukarya. *PLoS One*, *8*(7), Article e67221. https://doi.org/10.1371/journal.pone.0067221

Stevens, I. T., Irvine-Fynn, T. D. L., Edwards, A., Mitchell, A. C., Cook, J. M., Porter, P. R., Holt, T. O., Huss, M., Fettweis, X., Moorman, B. J., Sattler, B., & Hodson, A. J. (2022). Spatially consistent microbial

biomass and future cellular carbon release from melting Northern Hemisphere glacier surfaces. *Communications Earth & Environment*, *3*(1), 275. https://doi.org/10.1038/s43247-022-00609-0

Stibal, M., Gözdereliler, E., Cameron, K. A., Box, J. E., Stevens, I. T., Gokul, J. K., Schostag, M., Zarsky, J. D., Edwards, A., Irvine-Fynn, T. D. L., & Jacobsen, C. S. (2015). Microbial abundance in surface ice on the Greenland Ice Sheet [Original Research]. *Frontiers in Microbiology*, *6*. https://doi.org/10.3389/fmicb.2015.00225

Tedstone, A. J., Bamber, J. L., Cook, J. M., Williamson, C. J., Fettweis, X., Hodson, A. J., & Tranter, M. (2017). Dark ice dynamics of the south-west Greenland Ice Sheet. *Cryosphere*, *11*(6), 2491-2506. https://doi.org/10.5194/tc-11-2491-2017

Telling, J., Anesio, A. M., Tranter, M., Irvine-Fynn, T., Hodson, A., Butler, C., & Wadham, J. (2011). Nitrogen fixation on Arctic glaciers, Svalbard. *Journal of Geophysical Research: Biogeosciences*, *116*(G3). https://doi.org/https://doi.org/10.1029/2010JG001632

Telling, J., Stibal, M., Anesio, A. M., Tranter, M., Nias, I., Cook, J., Bellas, C., Lis, G., Wadham, J. L., Sole, A., Nienow, P., & Hodson, A. (2012). Microbial nitrogen cycling on the Greenland Ice Sheet. *Biogeosciences*, *9*(7), 2431-2442. https://doi.org/10.5194/bg-9-2431-2012

Tveit, A. T., Urich, T., Frenzel, P., & Svenning, M. M. (2015). Metabolic and trophic interactions modulate methane production by Arctic peat microbiota in response to warming. *Proceedings of the National Academy of Sciences*, *112*(19), E2507-E2516. https://doi.org/doi:10.1073/pnas.1420797112

Vandecraen, J., Chandler, M., Aertsen, A., & Van Houdt, R. (2017). The impact of insertion sequences on bacterial genome plasticity and adaptability. *Critical Reviews in Microbiology*, *43*(6), 709-730. https://doi.org/10.1080/1040841X.2017.1303661

Varliero, G., Rafiq, M., Singh, S., Summerfield, A., Sgouridis, F., Cowan, D. A., & Barker, G. (2021). Microbial characterisation and Cold-Adapted Predicted Protein (CAPP) database construction from the active layer of Greenland's permafrost. *FEMS Microbiology Ecology*, *97*(10). https://doi.org/10.1093/femsec/fiab127

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*, *27*(5), 737-746. https://doi.org/10.1101/gr.214270.116

Wang, B., Lin, A. E., Yuan, J., Novak, K. E., Koch, M. D., Wingreen, N. S., Adamson, B., & Gitai, Z. (2023). Single-cell massively-parallel multiplexed microbial sequencing (M3-seq) identifies rare bacterial populations and profiles phage infection. *Nature Microbiology*, *8*(10), 1846-1862. https://doi.org/10.1038/s41564-023-01462-3

Wensel, C. R., Pluznick, J. L., Salzberg, S. L., & Sears, C. L. (2022). Next-generation sequencing: insights to advance clinical investigations of the microbiome. *The Journal of Clinical Investigation*, *132*(7). https://doi.org/10.1172/JCI154944

Xia, Y., Li, X., Wu, Z., Nie, C., Cheng, Z., Sun, Y., Liu, L., & Zhang, T. (2023). Strategies and tools in illumina and nanopore-integrated metagenomic analysis of microbiome data. *iMeta*, *2*(1), e72. https://doi.org/https://doi.org/10.1002/imt2.72

Xue, Y., Lanzén, A., & Jonassen, I. (2020). Reconstructing ribosomal genes from large scale total RNA meta-transcriptomic data. *Bioinformatics*, *36*(11), 3365-3371. https://doi.org/10.1093/bioinformatics/btaa177

Yue, Y., Huang, H., Qi, Z., Dou, H.-M., Liu, X.-Y., Han, T.-F., Chen, Y., Song, X.-J., Zhang, Y.-H., & Tu, J. (2020). Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC Bioinformatics*, *21*(1), 334. https://doi.org/10.1186/s12859-020-03667-3