Project plan

Paper III: Metabolic Roles of Uncultivated Bacterioplankton Lineages in the Northern Gulf of Mexico "Dead Zone"

Background and aim of the project

Some marine regions have something called "dead zones" which is when the dissolved oxygen levels are very low – this can either be under a long time or occur seasonally. These dead zones have a negative impact on the ecology and economy since these low oxygen regions are unfit for many species, and thus understanding them better may be very valuable.

One of the main causes of these dead zones is microbial respiration. Marine emissions such as nitrogen-based fertilizer from agriculture can lead to algal blooms which then act as an excess source of carbon for the metabolism of aerobic microbes. These heightened levels of respiration utilize great amounts of oxygen, leading to a strong depletion in the ocean floor and thus the creation of dead zones. Previous knowledge on this topic has mainly been generated from studies on deep water, naturally occurring dead zones, which can be permanent due to a constant source of nutrients. However, a coastal dead zone that is over 20,000 square kilometers in size can be found on the continental shelf of the northern Gulf of Mexico due to fertilizer emissions in the Mississippi river from agriculture in the United States. Contrary to previous study areas with more of a permanent state of low oxygen, this dead zone is seasonal which opens for an interesting analysis on how the microbes in this area function during periods of low oxygen. More insight in their metabolism and gene expression may become valuable in alleviation plans of the situation in the northern Gulf of Mexico seasonal dead zone.

The questions to answers are "What metabolic expressions are present in these microbes during periods with low dissolved oxygen that can give clues to the seasonality of the dead zone?"

Analyses and software

This study will analyze reconstructed genomes of microbes in six samples from six different places of the northern Gulf of Mexico continental shelf. When analyzing the genomes, we will especially look for gene expressions related to metabolism in order to gain clues regarding the seasonality of the low levels of dissolved oxygen.

The sequencing method that is used in the study to generate the raw data is Illumina HiSeq 2000. The quality of the reads (both DNA and RNA) gained from the Illumina sequencing will be quality assessed using FASTQC. The RNA reads will then be trimmed using Trimmomatic and then be quality assessed again with FASTQC.

Thereafter Megahit will assemble the microbial genomes from the sequencing data. The data will be provided to Megahit in fastq format. This process will be very time-consuming and thus it should be submitted as a batch job overnight for efficiency.

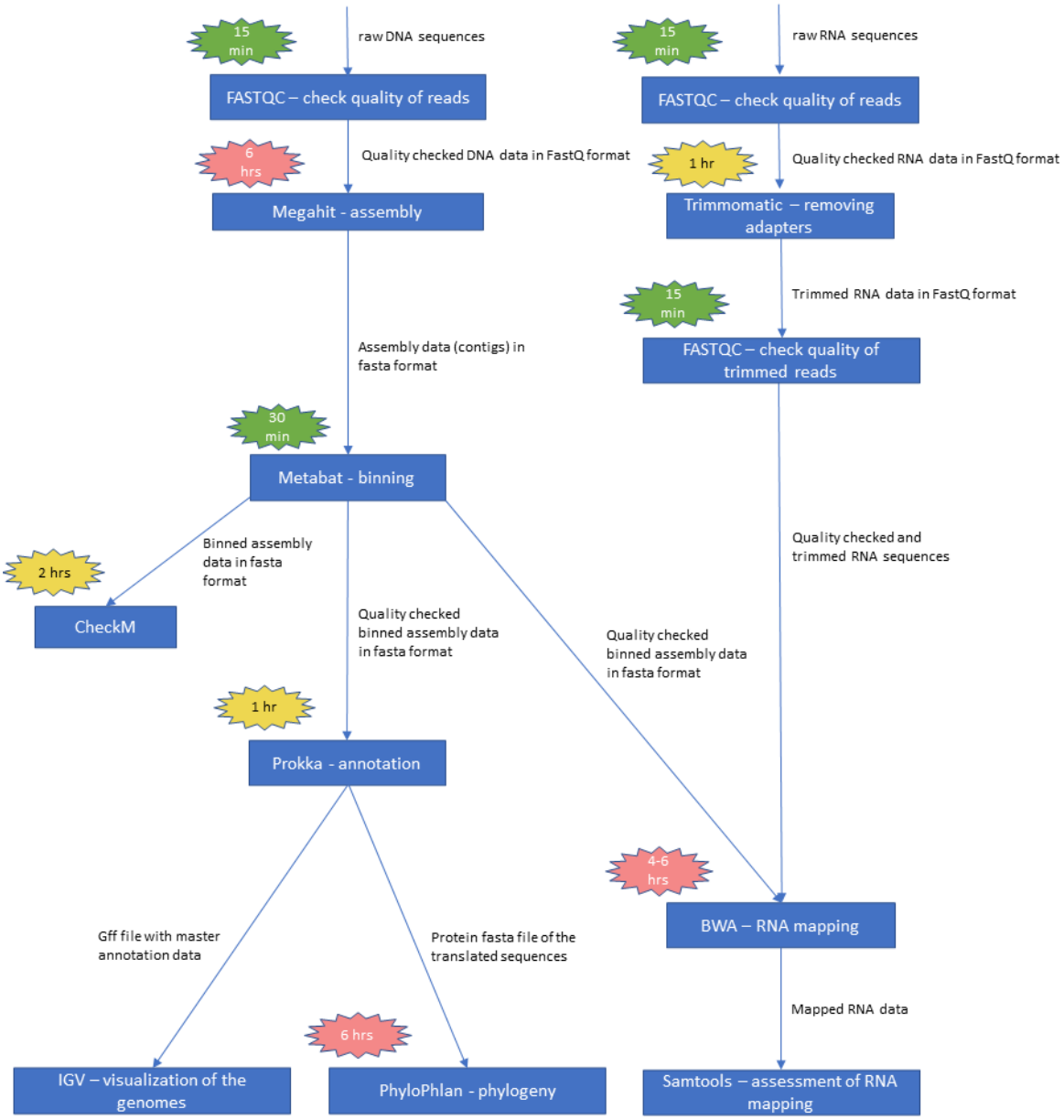
The assembly data will then be provided to Metabat for binning. The binning will sort the contigs from the assembly into different bins corresponding to different organisms. The

quality of the binning will be examined using CheckM. This process will require some time and thus it is preferable if it is submitted as a batch job.

The Prokka software will then use the binned and assembled genomic data to perform annotation. This will provide our genomes with information regarding gene functionality. The software IGV will be used to display the gff output file which is the master annotation file generated by Prokka.

Then RNA mapping will be performed using BWA. The RNA mapping will map the RNA sequence data to the reconstructed genomes. The input to BWA will be the reconstructed genomes from metabat and the trimmed RNA sequences. The mapping will be assessed using samtools. Here it is important to convert the files from SAM format to BAM format in order to save memory usage since SAM files are very large.

PhyloPhlan will use the translated protein sequences from Prokka and construct a phylogenetic tree.



Scheduling

In order to finish this project on time, these are the following deadline that I will try to follow.

2023-MAR-29: Finish project planning

2023-MAR-31: Examine raw data (both DNA and RNA) and quality check reads with FastQC. Submit batch job to assemble data with Megahit. Trim RNA adapters with Trimmomatic. Quality check RNA data again after trimming.

2023-APR-13: Examine results from Megahit.

2023-APR-14: Attend compulsory computer lab

2023-APR-18: Bin the assembly data with Metabat, submit a batch job for quality assessment via CheckM

2023-APR-19: Examine the results from CheckM. Annotate the quality checked data with Prokka. Prepare for RNA mapping and submit batch job for BWA.

2023-APR-25: Assess RNA mapping with Samtools + visualize genomes in IGV.

2023-APR-26: Continue assessing RNA mapping and visualization in IGV. Submit batch job for PhyloPhlan

2023-MAY-02: Attend compulsory computer lab

2023-MAY-10: Examine results from PhyloPhlan

2023-MAY-11: Analyze results + prepare presentation

2023-MAY-16: Analyze results + prepare presentation

2023-MAY-17: Presentation of results

By the compulsory computer lab at the second of May, I should be finished running all the software which should leave plenty of time for analysis and interpretation of results. The bottlenecks that I must be aware of are the times required for running Megahit, CheckM, BWA and PhyloPhlan, since these softwares take several hours to run. Therefore I will opt to submit these as batch jobs overnight so I won't spend any time waiting.

Data

The raw data that I will use is DNA and RNA data. These sequencing data were generated from samples in six different locations. The different samples were taken from different depths and have different amounts of dissolved oxygen. The data was sequenced via Illumina HiSeq 2000 and generated reads of 100 bp that is paired ended. The source of the DNA sequences is metagenomic while the source of the RNA sequences is metatranscriptomic. The size of the data from the different environmental sites is ~3GB-7.6GB for the DNA data and ~70Mb-28GB for RNA data. Since these datasets are very large, we will work with subsets in order to perform the project within the limited time frame and be able to present on May 17th.

Project Organization

In order to work efficiently I plan to keep my data organized from the beginning. The raw data is already downloaded and made available for us students, and I will use soft links to it in order to save space. Metadata for the sequencing data will be collected from public sources and stored in excel spreadsheet. I will always utilize clear and interpretable names for all my datafiles, so it is understandable for someone else than me. Separate folders will be created for analyses, code and data. Github will be used for the code since version control is very important. In every session, I will clean my working space and delete all files that I do not need, and make sure that everything is stored properly so that my working space can be easily navigated by someone else than me. Large files will be compressed, and SAM files will be converted to BAM files in order to minimize memory usage.