Genome Analysis 2021

Project Plan

Tomas Cumlin

The aim of the project

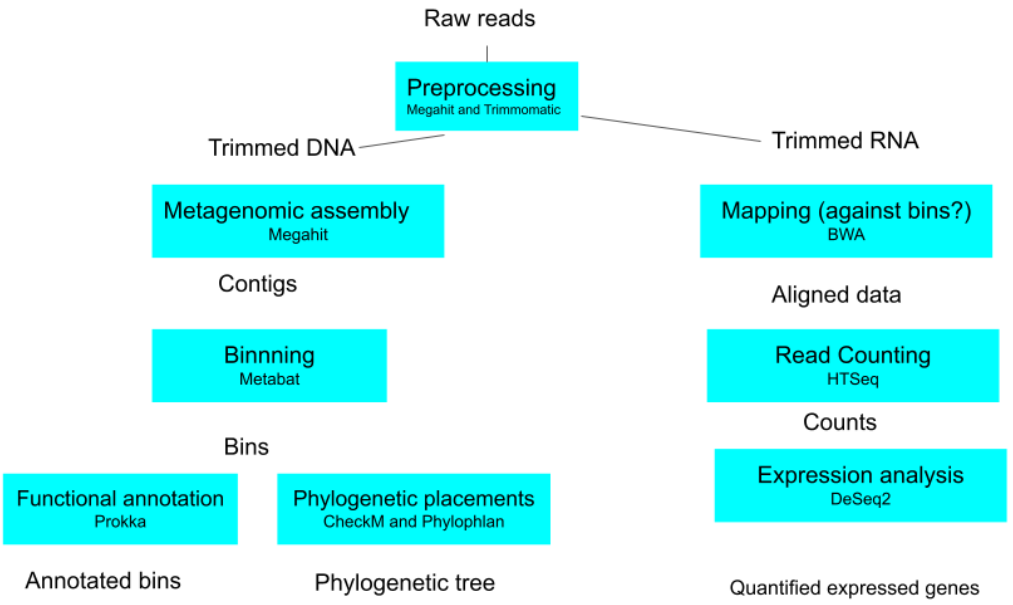
The aim of the project is to obtain the metabolic functions of some cosmopolitan bacterioplankton lineages abundant in the northern Gulf of Mexico. The purpose of this is to understand their ecological role within an aquatic environment that has low levels of dissolved oxygen. This is initially done by predicting which genes are expressed in the genomes of these species.

The major goals of this project are:

- To assemble the metagenomic data
- Find out which genome belongs to which species
- Functional annotation via metabolic reconstruction

Analysis steps

A summary of which analysis are necessary, in what order and the estimated time per step.



Summary of the project plan. It is not complete. It need quality checks.

Step	Analysis	Purpose	Software	Time
1	Reads preprocessing	Quality check	Trimmomatic and FastQC	15 min
2	Metagenomic assembly	Assemble RNA and DNA	Megahit	6 h
3	Assembly evaluation	See if assembly is succesful	Quast	45 min
4	Binning	Assign genomes to correct species	Metabat	>30 min
5	Quality check of assembly and bins	Quality evaluation	CheckM	2 h
6	Basic phylogenetic placement of bins	taxonomic ID	CheckM	2 h
7	Phylogenetic placement	Reconstruct phylogenetic tree	Phylophlan	6 h
8	Functional annotation	Metabolic reconstruction	Prokka	1 h
9a	Mapping	align the RNA against ref genome	BWA	4-6 h
9b	Read counting	count the aligned reads	HTSeq package in Python	n/a
9c	Expression analysis	which genes are more expressed	DeSeq2 package in R	n/a

Step	Extra Analysis	Purpose	Software	Time
9a	Abundance of different organisms/bins	n/a	BWA	n/a
7	Refine taxonomic ID	n/a	FastTree2	n/a
4 (after binning)	Metabolic pathway reconstructions	n/a	Submit to IMG (Integrated Microbial Genomes)	n/a
after 9	Analysis of expression data	n/a	n/a	n/a
after 9	Comparisons across bins	n/a	n/a	n/a
after 9	Comparative genomics of bins	n/a	n/a	n/a
9c	Ortholog gene clustering of bins	n/a	DeSeq2 package in R	n/a

* Yellow field means that the time step is extra time-consuming

The Time Frame

* Not completed

9/4: Datalab

14/4: Datalab

15/4: Datalab

20/4: Datalab

23/4: Datalab

28/4: Datalab

29/4: Datalab

03/5: Datalab

04/5: Datalab

11/5: Datalab

17/5: Datalab

19/5: Datalab

25/5: Presentation

The Data

*DNA and RNA which have been sequenced using Illumina HiSeq 2000.

DNA

SAMN05791315: 3 GB
SAMN05791316: 3.5 GB
SAMN05791317: 6.3 GB
SAMN05791318: 6.5 GB
SAMN05791319: 12.8 GB
SAMN05791320: 3.3 GB

Total: 35.4 GB

RNA

SAMN05791321: 178.1 Mb
SAMN05791322: 630 Mb
SAMN05791323: 164.6 Mb
SAMN05791324: 71.3 Mb
SAMN05791325: 28.8 Gb
SAMN057913256: 1.7 Gb

Total: 31.5444 GB

Whereof Uppmax stores up to 32 GB.

Project Organization

* I have not done a map on how to organize the folders more precisely yet.
In my project directory, there will be a data folder for my raw DNA-and RNA data. There will be another folder where I save all my code. There will be a folder where I have all my analysis. Finally, I will have a folder where I store my results, images and other information related to this project.