METAGENOMICS PROJECT

Daniel Martín, Guillem Miró, Marc Pinós

2023-11-13

INTRODUCTION

Sequencing and metagenomics are biological disciplines that study the genetic and epigenetic sequence information of organisms. By sequencing their genome they attempt to understand several genetic concepts such as, genetic functionality, genotype, phenotype, and inheritance among others.

In this project a metagenomic analysis will be performed using an initial sequenced data set of an unknown bacterial species as reference. By performing an exhaustive methodology, latter sequence will be sorted out in order to identify its origin and produce several analysis.

Objectives

- To perform a metagenomic analysis from a genomic sequence of unknown origin.
- To select relevant and higher quality information from the sequence in order to properly work with it.
- To identify the taxonomic classification of the unknown sequence.
- To infere functional categories of the genomic protein sequences and analyze a selected category.
- To produce a phylogeny of the reference genome and its nearer species.

METHODS

Practical 4: Assembly of a bacterial genome

Raw sequence data will be assembled and analyzed so as to purge for a higher quality sample. In our particular case, sequencing data is composed in a FASTQ format, which provides the sequence itself and its quality information.

1. Read Quality Control

The sequence data will be analyzed by *Assess Read Quality with FastQC* so as to make en exhaustive report of the reads' quality. This will enable the detection of possible noise, unwanted/impure fragments, sequencing mistakes, and base distribution rates.

**Parameters:* - Basic Statistics - Per base sequence quality - Per tile sequence quality - Per sequence quality scores - Per base sequence content - Per sequence GC content - Per base N content - Sequence length distribution - Sequence duplication levels

2. Trimming

Once all impure fragments are detected in the reads, trimming will be performed. By using *Trim Reads with Trimmomatic* an input read will be cut following a base pair (bp) threshold. For instance, starting and end point can be stipulated by selecting crop reads, which selects the number of bp to keep from start of the read.

The program will produce a new file which will correspond to the new sequence already trimmed.

3. Read Quality Control

The trimmed sequence will be assessed a new quality control (using *Assess Read Quality with FastQC* again) so as to identify all the compensations of quality compared to the raw analysis.

If all the previously-identified mistakes are compensated by the trimming, the analysis will be continued. The second quality control were assessed and showed a high quality score in all base pair position of all short reads, even in the head and tail fragments of each read. The result is shown in Figure XXX.

4. Assemble Read

All the reads will be stored into continuous sequences. In our case, *MEGAHIT* will be used. Compared to other softwares it is believed to be faster. All the reads will be conjugated to create long DNA fragments, known as *contigs* (assemble continuous genome fragments).

5. Binning

Previous *contigs* will be grouped in several *bins*. The term bin corresponds to a putative population genomes created by a selection of properties, such as codon usage, genome related statistics, GC proportion, read length, among other parameters. The analysis will be performed using *MaxBin2*.

6. Bin quality control

Finally, all bins created will be processed under a quality control. It will provide an estimate of genome completeness and contamination by plotting each bin compared to its expected distribution of a typical genome. *CheckM* will be used to perform the control.

In particular, we will look for high quality genomes, which correspond to >90% completeness and <5% of contamination.