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 infer 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.

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.

Practical 5: Bacterial genome identification

Once all quality reads are selected, the analysis must begin. In particular, a taxonomic identification of out unknown sample will be performed.

In the case of bacteria, as they do not reproduce sexually, species classification is induced by the computation of ANI (average nucleotide identity). Bacteria are known to exchange core genes between mainly same species, and being unusual (or at least more difficult) to pass to different ones. Therefore, computing percentage of identity between a reference species and the sample will enable an identity classification:

gANI:

- 95% same species
- 85% same genus
- 70% same family

1. Genome annotation

The higher quality bin from practical 4 will be selected to perform the analysis. *PROKKA* software will be used to annotate the sample genome (placed in the bin) in order to identify several pieces or regions prone to be compared.

2. Genome set

Selecting latter annotated genome, we will create a genome set, enabilg a further research with the sample.

3. Microbe classification

The genome set will be taxonomically classified using *GTDB*. As an output a taxonomical tree will promote the identification of the bacterial sepecies present in our sample.

4. GTDB: genome taxonomy data base

GTDB will be used to download the *Genebank assembly accession* of out bacteria and acquire a FASTA reference sequence.

5. ANI computation

Latter FASTA sequence (type strain genome) and the annotated genome (Meta genome assembled genome) from step 1 will be compared using *FastANI*, which will compute the ANI.

In particular, an accurate ANI will be estimated by the comparison of nucleotide of orthologous gene pairs shared between the 2 sequences given.

Practical 6: Genome annotation I

In practical 6 we will proceed with the genome annotation of both sample and refernce genome.

1. Reference genome annotation

Following last practical's procedure, the genome extracted from GTDB will be also annotated.

2. Data exploration and functional categories

By selecting the annotated sample, we will select the binocular option in order to explore all data gathered.

Thanks to *PROKKA annotation pipeline* we will be able to acknowledge provenance, linked samples, genome overview, taxonomy, publications, and assembly and annotation from our sample. The latter category will be selected to find the functional categories encoded in our genome, which corresponds to a set of proteins prone to be expressed by our bacterial specimen.

The functional categories will be compared to the reference genome estracted from GTDB.

Afterwards, we will download a GFF (*gene feature format*) and a FASTA files to work on both genomes. A specific functional category will be selected and we will compute:

- The number of tRNAs of our sample.
- The number of rRNAs of our sample.
- The number of genes encoding proteins without known function.

Finally, using *RStudio* we will create a barplot of the types of genes found in our MAG and our genome of reference (type strain).

Practical 7: Genome annotation II

1. Multiple sequence alignment

Both nucleotide and amino acid sequences from our MAG and the selected reference genome will be aligned with MAFFT (Multiple Alignment using Fast Fourier Transform). The resulting file containing the alignment of both sequences will be visualized with Jalview.

2. Pairwise sequence alignment

In the same way as in the previous step, the pairwise alignment socre for both nucleotide and amino acid sequences from our MAG and the reference genome will be assessed using SMS (Sequence Manipulation Suite) from *bioinformatics.org*. This parameter takes into account the number of matches (i.e. A-A) and missmatches (i.e. A-T) between both sequences (as well as the possible gaps present in one of them) and computes them in order to give a value referring to their alignment level.

NO ENTENC LO DE LA ÚLTIMA PART DE LES SCORING MATRIX

Practical 8

RESULTS AND DISCUSSION

Practical 4

Upon acquiring the raw data for analysis, we conducted a thorough quality control assessment of all base pair positions within the short reads. Remarkably, we observed consistently high-quality scores across the entire span of the reads, including both head and tail fragments as depicted in *Figure 1.* Consequently, a decision was made to forgo Trimming procedures, aiming to preserve information from the extremities of the short reads.

Following genome assembly, seven distinct bins were generated (*Figure 2*). Notably, bins 01, 03, and 04 exhibited completeness levels nearing maximum values, approximating 100%. This indicates the presence

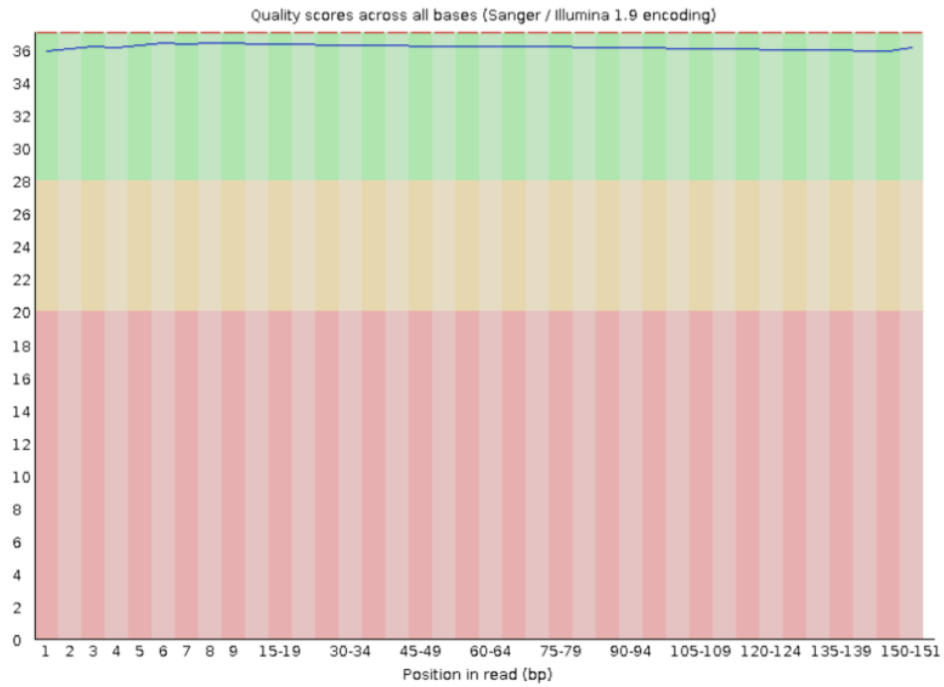

Figure 1: Quality check reads

of all anticipated genes within the Metagenome-Assembled Genomes (MAGs). Additionally, these bins demonstrated a contamination level below 1%, affirming the appropriateness of the assembly process due to the initial high-quality short reads, enabling seamless assembly (*Figure 3*). Thus, bins 01, 03, and 04 can be considered of high quality, attaining >50% completeness and <5% contamination.

Upon analyzing the distribution of the short reads utilized for MAG assembly, it was observed that these ranged between 145-152 base pairs, slightly longer than the optimal size for assembly. Furthermore, approximately 5% of the sequences were identified as being repeated more than 10 times. Considering these findings, the GC content of our MAG might slightly deviate from the theoretical distribution (*Figure 4*).

Practical 5

We opted to focus on bin03 due to its exceptional completeness level (100%) and minimal contamination (0.91%). Following the annotation of our MAG and subsequent genome set generation, we employed GTDB classification to determine the bacterial strain origin of this genome. The outcomes are presented in *Figure 5*, indicating a match between bin03 and *Bifidobacterium dentium* with an Average Nucleotide Identity (ANI) score of 98.73, as illustrated in the phylogenetic tree displayed in *Figure 6*.

Subsequently, after obtaining the reference genome for *B. dentium*, we conducted a comparative analysis with our MAG, computing the nucleotide identity between both sequences. This assessment yielded an ANI distance estimate of 98.72% (*Figure 7*), indicating a similarity of 98.72% in nucleotide composition between the sequences. The comparison details can be visualized in *Figure 8*.

Practical 6

If we compare the number of features of each genome (Bin03 and reference) we have found 2216 in our genome in comparison to 2137 of the reference genome. From these features, the functional categories codified in both genomes are shown in *Figure 9*. We can see that more or less in both genomes we find the same genes in

each category in exeption of: a) Stress response, b) Cell wall and capsule, c) RNA metabolism, d) Membrane transport and e) DNA metabolism.

On the other hand, it is necessary to find how many tRNA's, rRNA's and hypothetical coding genes are found in our MAG and in the reference.

Upon comparing the number of features present in each genome (Bin03 and the reference), we identified 2216 features in our genome compared to 2137 in the reference genome. Among these features, the functional categories encoded in both genomes are illustrated in *Figure 9*. It is apparent that both genomes harbor similar genes within each category, except for: a) Stress response, b) Cell wall and capsule, c) RNA metabolism, d) Membrane transport, and e) DNA metabolism.

Additionally, it is pertinent to ascertain the counts of tRNAs, rRNAs, and hypothetical coding genes present in our MAG compared to the reference.

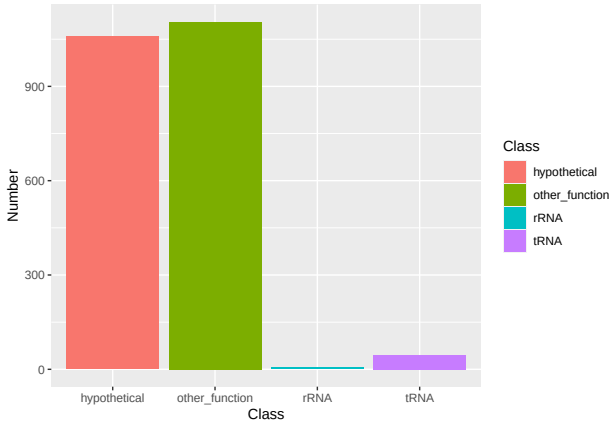

Figure 2: MAG graphic

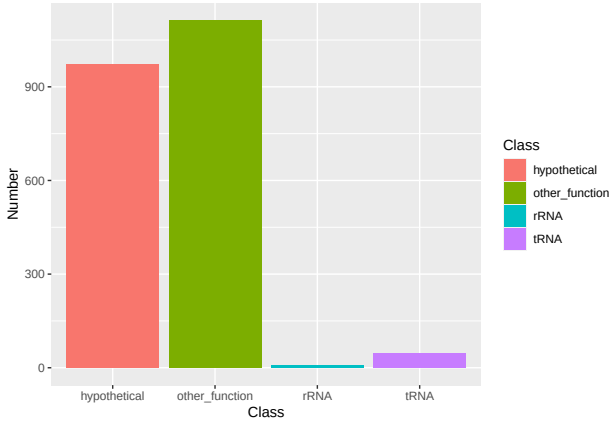

Figure 3: Reference Graphic

Practical 7

Practical 8

CONCLUSIONS

REFERENCES