UNIVERSITÀ DI TRENTO
Dipartimento di Biologia Cellulare, Computazionale e Integrata
CIBIO

Master Degree in
Quantitative and Computational Biology

Project group
Computational Microbes Genomics

Group:
Andrea Tonina, Lorenzo Santarelli, Gloria Lugoboni

Date 12/04/2023

Contents

1 Introduction

1.1 Metagenome Sequencing, Assembly, and Binning

Metagenome sequencing enables the construction of metagenomes-assembled genomes (MAGs). A MAG can be seen as a microbial genome obtained by a preliminary passage of genome assembly of high quality contigs. This kind of analysis enables us to identify novel species thanks to a passage of annotation and taxonomic classification [1].

A typical metagenome project involves a specific pipeline, a step of sample processing and sequencing, a step of assembly and finally a step of binning followed by genome-annotation. This whole process is than completed with a statistical analysis [2].

Metagenomics is possible thanks to the study of DNA genomes, the sequencing is possible using a variety of novel sequencing technologies and platforms like Roche 454 sequencing, Illumina sequencing, and ion torrent Personal Genome Machine (PGM) [3].

Thanks to the proccess of assembly it is possible to reconstruct genomes. This method is based on a process of aligment and merging of overlapping sequences, creating large contiguous regions (contigs) [4].

After the process of assembly is completed, contigs are grouped by their organism of origin into bins, using a process known as binning [5]. The selection of high quality bins enables the identification of MAGs, these are characterized by a high completness and low levels of contamination and are used to operate taxonomic annotation and gene prediction [6]. These can be grouped together in the same species genome bin (SGB) if they exceed a certain threshold of nucleotide identity, with a treshold of the 5% for genomic identity. It is possible to assign a taxonomic label based on the presence (or not) of characterized genomess [7]. If a genome with associated taxonomy is not availabe, we talk about known SGB (kSGB), while in the opposite case, we talk about unknown clades (uSGB) [8].

DA SPIEGARE IL PANGENOME ANAYSIS E LE ALTRE ANALISI CHE FACCIAMO??

PANGENOME: A bacterial species can be described by its pan-genome, which is composed of a ‘core genome’ containing genes present in all strains, and a ’dispensable genome’ containing genes present in two or more strains and genes unique to single strains. Given that the number of unique genes is vast, the pan-genome of a bacterial species might be orders of magnitude larger than any single genome [9].

1.2 uSGB 15132

We were provided with a set of 30 high-quality prebinned metagenomes grouped in the same uSGB labelled SGB15132.

The bins have a completeness higher that 97.3, as shown in the Figure 1.1 and the maximum redundacy registered is equal to 2.25.

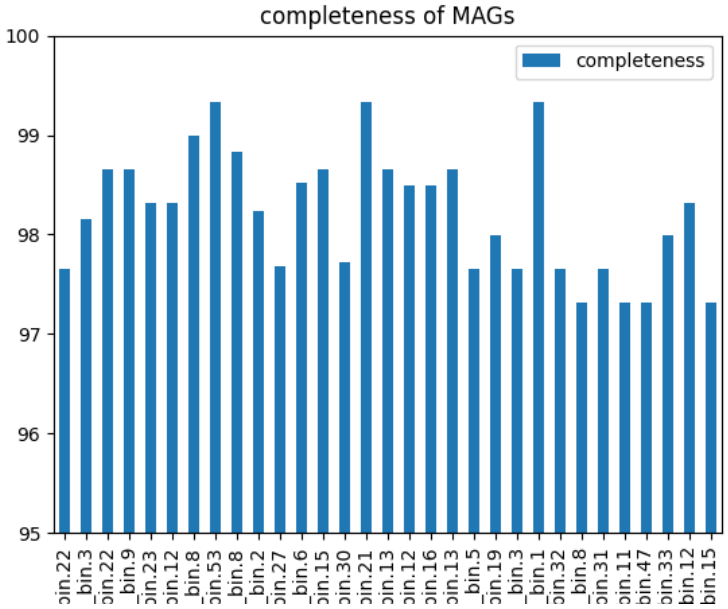

Figure 1.1: Completeness distribution of the given MAGs.

2 Methods

2.1 Softwares and parameters used

- What software you used for each purpose, what parameters

2.1.1 Genome annotation (Prokka)

Prokka is a fast and accurate command line software tool used to annotate prokaryoyic genomes. It produces standards-compliant output files that can be used for further analysis or viewing in genome browsers.

Prokka expect one single input file in a FASTA format, containing an assembled genome. The process of annotation is possible thanks to the comparison of the gene codes with a large database of known sequences, identifying the best match as the most significative one and therefore associating the labelling and the relevant features to the gene codes. Prokka use this method in an hieratical manner, using initially small and reliable databases moving only at the end of the process to protein family databases. Prokka produces several outputs file, listed in the Figure 2.1 [10].

We need to specify several parameters, specifically, the input files, our MAGs, the output directory `--outdir` and the parameter `--kingdom Bacteria` that is needed to specify the annotation mode, to make the prokka more fast.
`prokka --kingdom Bacteria --outdir SGB15132_prokka_output .f*na` .

Suffix	Description of file contents
.fna	FASTA file of original input contigs (nucleotide)
.faa	FASTA file of translated coding genes (protein)
.ffn	FASTA file of all genomic features (nucleotide)
.fsa	Contig sequences for submission (nucleotide)
.tbl	Feature table for submission
.sqn	Sequin editable file for submission
.gbk	Genbank file containing sequences and annotations
.gff	GFF v3 file containing sequences and annotations
.log	Log file of Prokka processing output
.txt	Annotation summary statistics

Figure 2.1: Prokka outputs files [10].

2.1.2 Pangenome analysis (Roary)

Roary is a tool that enables the construction of large-scale prokaryote pangenomes, identifying the core and accessory genes.

The input file to Roary is a GFF file containing sequences features.

Roary collects the coding regions from the annotated input genome. It operates a clustering process creating a network and defining a phylogenetic tree. A matrix is therefore obtained and the pangenome (core genes and accessory genes) is defined. The process of clustering is based on the minimum percentage of identity, setted to 95% by default [11].

Roary returns three graphs, the newick tree associated to the pamgenome table, a pie chart of the breakdown of genes and the number of isolate they are present in, a graph with the frequency of genes versus the number of genomes. [12]

There are some main parameters that need to be specified to roary, specifically, the input `.gff` files; the output directory `-f roary_out` ; the -i parameter, specifing the percentage identity of blastp, here used at 95%, `-i 95` ; the -cd parameter, percentage of isolates a gene must be in to be considered part of the core genome, here setted at 90%, `-cd 90` ; the `-e` parameter, to perform a core gene alignment; the `-n` parameter, to use

mafft as the tool for the multiple sequence alignmet, making the process faster and finally the parameter `-p`, needed to specify the number of threads, increasing therefore the speed [11].
`roary .gff -i 95 -cd 90 -e -n -p 8`.

2.1.3 Taxonomic assignment (PhyloPhlAn 3.0)

PhyloPhlAn 3.0 is an accurate and rapid tool to perform microbial genome characterization and phylogenetic analysis both of newly assembled microbial genomes and metagenomes. PhyloPhlAn 3.0 can integrate public genome resources/information to the genomes in input and is also accurate at the strain and species level. [13] The main input to be sopecified are :...

2.1.4 Phylogenetic analysis (Roary+FastTree)

Roary enable us to generate a core gene alignment of our uSGB using as specific parametres -e, –mafft and -p, **roary -e –mafft -p 8 *.gff**. This aligment can be used to construct a phylogenetic tree, this is possible using FastTree, a tool for constructing large phylogenies, estimating their reliability. FastTree exploit Neighbor-Joining and nearest neighbor interchanges to create a phylogentic tree. [14]

2.1.5 Association with host data

3 Results and discussion

- Description of the set of bins: where do your MAGs come from, to what SGB do they belong, completeness and contamination

3.1 Genome annotation

After the gene annotation process, possible using prokka, we were able to identify that the number of the CDS (protein coding sequence) is slightly variable, spanning from a minimun value of 2651 to a max of 3935. For each MAG, more or less a half of the CDS are known proteins, while the other half is represented by RNA or hypotetical proteins.

3.2 Pangenome analysis

Each SGB strain was found to contain an average of 1317 genes that are present in every strain (core genome), plus 8407 genes that are absent in more than one strains (strain ¡30) (accessory genome). The accessory genes are also divided into genes present in only one strain (cloud genome) or genes present in two or more strains but not all strains (shell genome) [9]. The figure 3.1 show the divisio of the pangenome?.

what's the size of your pangenome? Is it closed or open? How many core and accessory genes?

3.3 Phylogenetic analysis and association with host data

comparison of phylogenetic trees based on accessory gene presence/absence or on core gene alignment. Do you detect clusters of strains? How do they associate with the metadata?

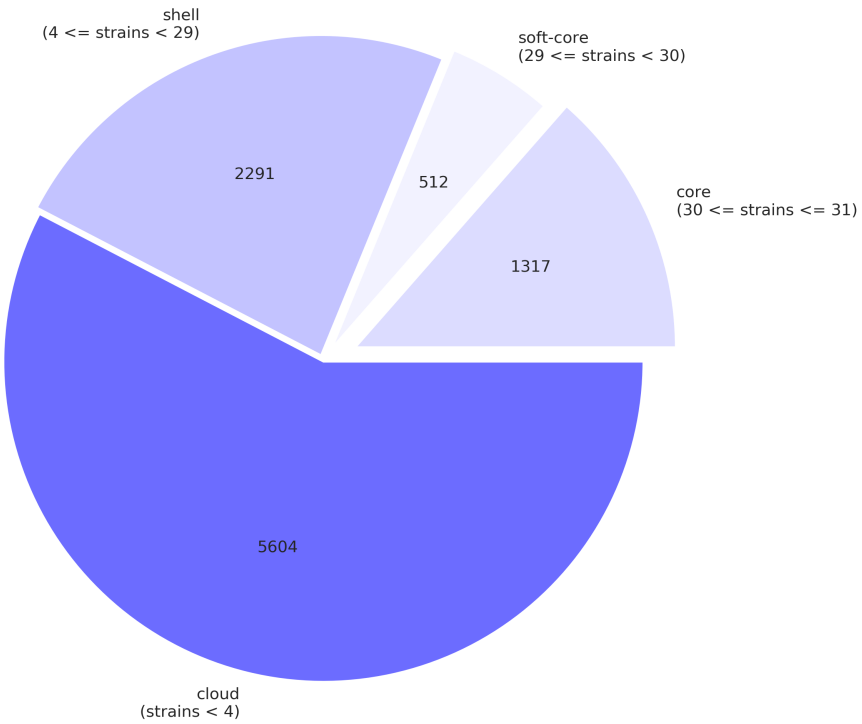

Figure 3.1: Pie chart showing the disvision of the pangenome

4 Conclusion

Bibliography

[1] Yang, Chao et al. *A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data,* Computational and structural biotechnology journal vol. 19 6301-6314. 23 Nov. 2021,
doi:10.1016/j.csbj.2021.11.028

[2] Thomas T, Gilbert J, Meyer F. *Metagenomics - a guide from sampling to data analysis.* Microb Inform Exp. 2012 Feb 9;2(1):3.
doi: 10.1186/2042-5783-2-3. PMID: 22587947; PMCID: PMC3351745.

[3] Reuter JA, Spacek DV, Snyder MP. *High-throughput sequencing technologies.* Mol Cell. 2015 May 21;58(4):586-97.
doi: 10.1016/j.molcel.2015.05.004. PMID: 26000844; PMCID: PMC4494749

[4] Churko JM, Mantalas GL, Snyder MP, Wu JC. *Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases.* Circ Res. 2013 Jun 7;112(12):1613-23.
doi: 10.1161/CIRCRESAHA.113.300939.

[5] Nissen, J.N., Johansen, J., Allesøe, R.L. et al. *Improved metagenome binning and assembly using deep variational autoencoders.* Nat Biotechnol 39, 555–560 (2021).
https://doi.org/10.1038/s41587-020-00777-4

[6] Chao Yang, Debajyoti Chowdhury, Zhenmiao Zhang, William K. Cheung, Aiping Lu, Zhaoxiang Bian, Lu Zhang, *A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data,* Computational and Structural Biotechnology Journal, Volume 19, 2021, Pages 6301-6314, ISSN 2001-0370,
https://doi.org/10.1016/j.csbj.2021.11.028.

[7] Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N. *Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle.* Cell. 2019 Jan 24;176(3):649-662.e20.
doi: 10.1016/j.cell.2019.01.001.

[8] Blanco-Míguez, A., Beghini, F., Cumbo, F. et al. *Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4.* Nat Biotechnol (2023).
https://doi.org/10.1038/s41587-023-01688-w

[9] Medini, Duccio et al. *The microbial pan-genome.* Current opinion in genetics and development vol. 15,6 (2005): 589-94.
doi:10.1016/j.gde.2005.09.006

[10] Torsten Seemann,*Prokka: rapid prokaryotic genome annotation,* Bioinformatics, Volume 30, Issue 14, July 2014, Pages 2068–2069,
https://doi.org/10.1093/bioinformatics/btu153

[11] Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. *Roary: rapid large-scale prokaryote pan genome analysis.* Bioinformatics. 2015 Nov 15; 31(22):3691-3.
doi: 10.1093/bioinformatics/btv421

[12] *https://sanger-pathogens.github.io/Roary/*

[13] Asnicar, F., Thomas, A.M., Beghini, F. et al. *Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0.* Nat Commun 11, 2500 (2020).
https://doi.org/10.1038/s41467-020-16366-7

[14] Morgan N. Price, Paramvir S. Dehal, Adam P. Arkin, *FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix, Molecular Biology and Evolution,* Volume 26, Issue 7, July 2009, Pages 1641–1650,
https://doi.org/10.1093/molbev/msp077