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 Motivation

To obtain information about a uSGB, starting from a set of 30 HQ bins (i.e. MAGs) -¿ Get the most out of a set of MAGs belonging to a uSGBs:

1.2 uSGB numero_usgb

Which uSGB are you working with, what is your focus

2 Methods

2.1 Softwares and parameters used

- What software you used for each purpose, what parameters 1-¿ introduzione sulla bash?

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 [1].

The main input to be specified are :...

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 [1].

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. [2]

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. [3]

DA AGGIUNGERE ANALISI DEL PANGENOME SE CLOSED O OPEN The main input to be sopecified are :...

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. [4] 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. [5]

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

what functions are encoded in your MAGs?
Hypothetical/annotated proteins

3.2 Pangenome analysis

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?

4 Conclusion

Bibliography

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

[2] 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

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

[4] 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

[5] 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