UNIVERSITÀ DI TRENTO
Dipartimento di
Biologia Cellulare, Computazionale e Integrata

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** **5**
1.1 Metagenome Sequencing, Assembly, and Binning . . . 5

2 **Methods** **7**
2.1 Softwares and parameters used . . . 7
2.1.1 Genome annotation (Prokka) . . . 7
2.1.2 Pangenome analysis and Phylogenetic analysis (Roary, Roary + FastTree) . . . 7
2.1.3 Taxonomic assignment (PhyloPhlAn 3.0) . . . 8

3 **Results and discussion** **9**
3.1 uSGB 15132 . . . 9
3.2 Genome annotation . . . 9
3.3 Pangenome analysis . . . 9
3.4 Phylogenetic analysis and association with host data . . . 10

4 **Conclusion** **13**

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].

With the term pangenome, we indicate the union of the *core genome*, containing genes present in all strains, and the *dispensable genome*, also called *accessory genome*, containing genes present in two or more strains and genes unique to single strains [9].

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

2 Methods

2.1 Softwares and parameters used

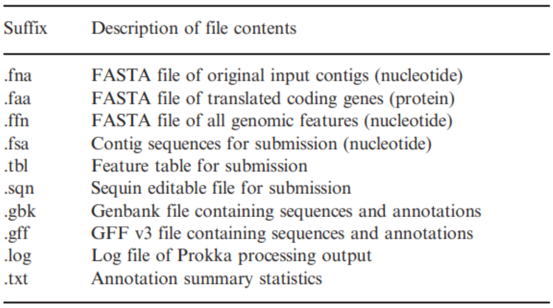
2.1.1 Genome annotation (Prokka)

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*` .



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 and Phylogenetic analysis (Roary, Roary + FastTree)

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].

The main output of Roary is a tree obtained using the presence and absence of the accessory genes. It is a tree used to have an initial insight of the data, grouping in a quick way the genomes based on their accessory genes [11]. It can be visualized using iTOL, an online tool for phylogenetic tree display [12].

Roary returns three graphs, the newick tree associated to the pangenome 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. [13]

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%; the , `-cd` parameter, percentage of isolates a gene must be in to be considered part of the core genome, here setted at

90%.

```
roary .gff -i 95 -cd 90
```

With Roary it is also possible to perform a core gene alignment to generate a more reliable tree. A core-genome alignment is more scalable with respect to the whole-genome alignment. It is an alignment useful to identify the core genes conserved in all aligned genomes and that can be very useful to infer the phylogeny [14].

The main parameters sto be specified are the `-e` parameter,needed 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
```
.

The core gene 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. [15] Specifically, we used FastTreeMP, that allow the parallelization of the steps needed in computing a tree [16]. The tree is obtained using the following code,

```
FastTreeMP -gtr -nt -out core_gene.tre core_gene_alignment.aln
```
,

in which the parameter `-gtr` express the generalized time-reversible model (to be used with nucleotide alignments only) while the parameter `-nt` is used to specify that the alignment is performed on nucleotides.

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 and allow the assign to each bin obtained via metagenomic assembly its closest species-level genome bins [17]. There are some main parameters needed to be specified, mainly, the input folder with the `-i` parameter; the output folder, with the parameter `-o`; the `--nproc` parameter, used to specify the CPUs that can be used; the `-n` parameter that allow us to decide how many SGBs (sorted by increasing average genomic distance) will be reported for each input bin in the output file; the `--database_update` parameter to update the databases file, the `-d` parameter to specify the name of the output database adn finally, the `--verbose` parameter to print to the bash [18]. The final command is the following:

```
phylophlan_metagenomic -i phylo -o phylo_out --nproc 4 -n 1 --database_update -d CMG2324 --verbose
```

3 Results and discussion

3.1 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 3.1 and the maximum redundacy registered is equal to 2.25.

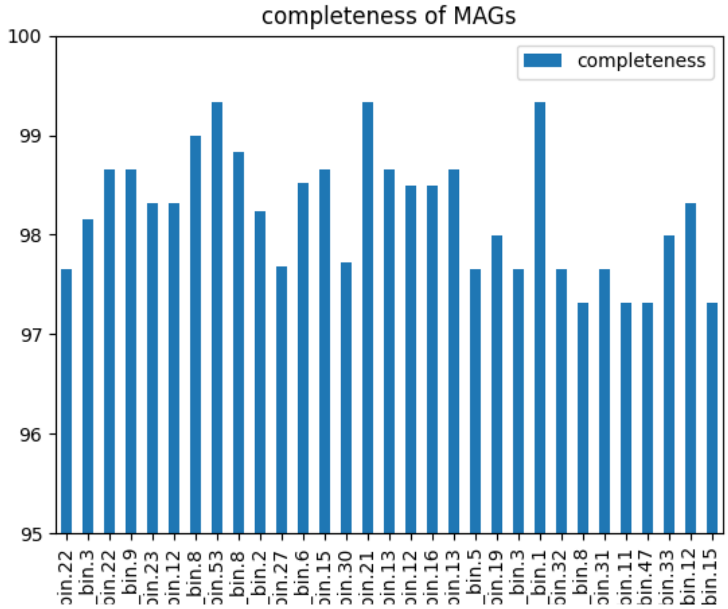


Figure 3.1: Completeness distribution of the given MAGs.

3.2 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, as shown in the Figure 3.2 more or less a half of the CDS are known proteins, while the other half is represented by RNA or hypotetical proteins.

3.3 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). This can be seen in the Figure 3.3, showing the frequency plot of the genes per genome, this plot gives a general overview of the frequency of genes within a whole genome set, typically these plots have an shape and most genes can be detected in a single genome or in all genomes [19]. Here it is shown that the number of genes present in all the genomes correspond to the number of core genes. 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 3.4 show the subparts constituting the pangenome, associated with the total number of genes.

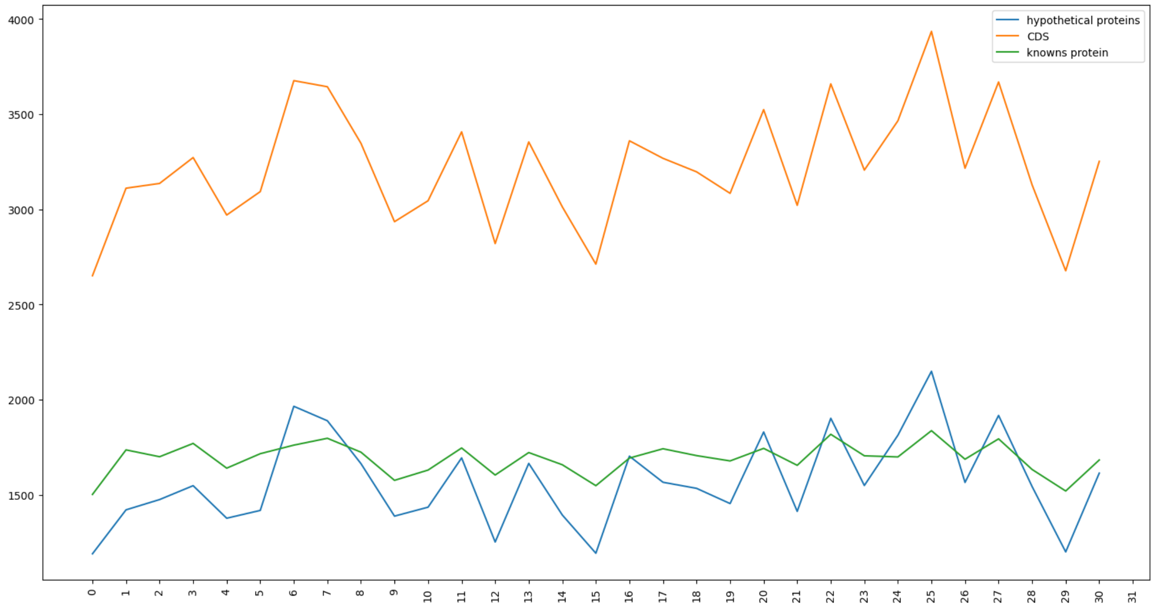


Figure 3.2: Main annotation results

Looking at the Figures 3.5 and 3.6 we can obtain/derive an idea on the pangenome, specifically if it is open or closed. It is important to remember that these results were obtained using only 30 metagenomes, making the process of discussion more complex. Even though, looking at the Conserved vs Total genes plot in Figure 3.5, we can argue that the total genes initial slope is very high but at the end, when almost all the genomes were added, the slope start to decrease. The conserved gene line is almost stationary, with litte changes when the genomes are all added. Only looking at this graph it is difficult to define if the pangenome is open or closed, therefore, the observation of the Unique vs New genes plot can be useful. Looking at the Figure 3.6, we can see that, adding new genomes doesn't give new information, the number of new genes is almost at zero when we add the last genome. This help us to state with a good security that the pangenome is closed.

3.4 Phylogenetic analysis and association with host data

PhyloPhlAn enable us to operate a phylogenetic analysis. The main output correspond to a list of the closest SGBs sorted by their increasing average sequence distance (Mash distance) in a tab-separated file [18]. The information of each SGB contains specific columns:

`my_bin` : is the input bin name

`(k|u)SGB_ID` : the SGB ID, k indicate a known SGB, u indicate an unknown SGB

`taxa_level` : taxonomic level the SGB has been assigned to

`taxonomy` : the full taxonomic label assigned to the SGB

`average_mash_distance` : the distance is calculated with respect to all the genomes in the SGB.

In our case, for each MAG we get the same taxa_level, the species Flavonifractor plautii, a strictly anaerobic rod shaped bacterium. It is a commensal of the human intestinal microbiota which is [20]

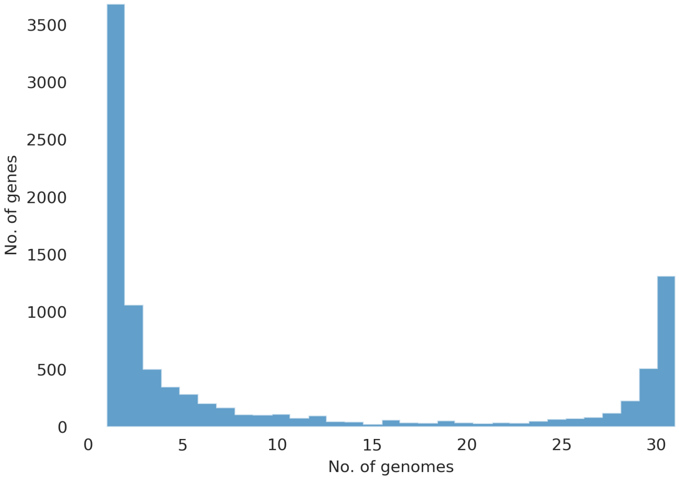


Figure 3.3: Pangenome Frequency

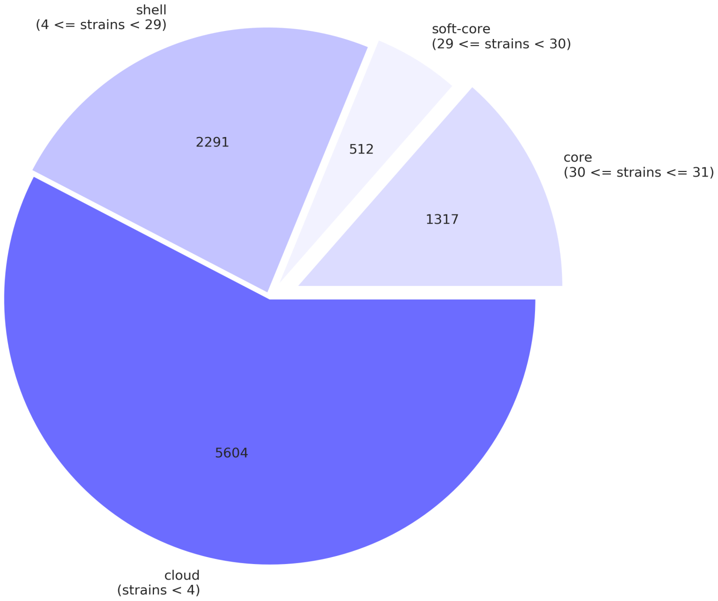


Figure 3.4: Core and accessory genome

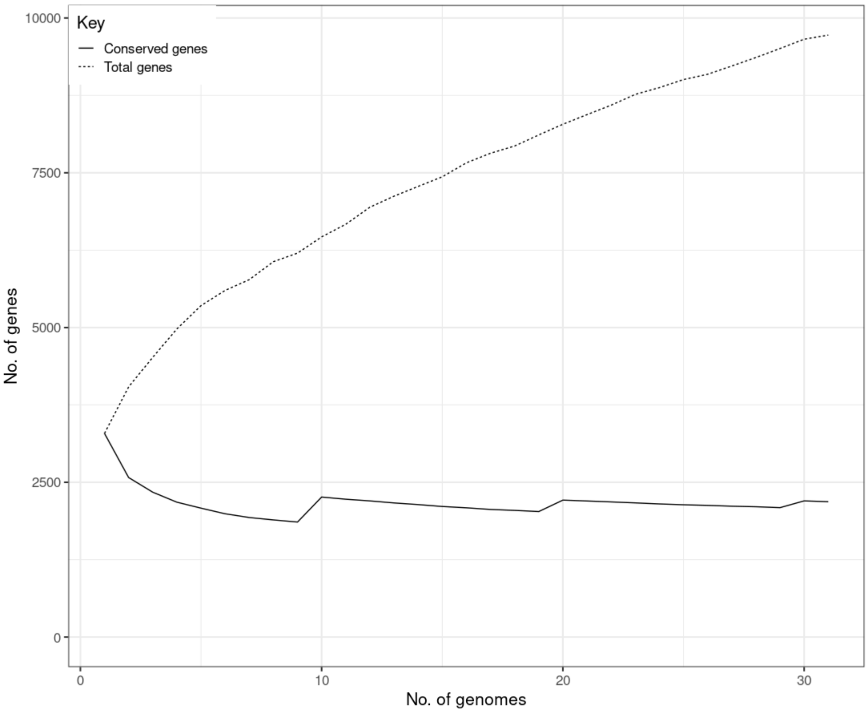


Figure 3.5: Conserved vs Total genes

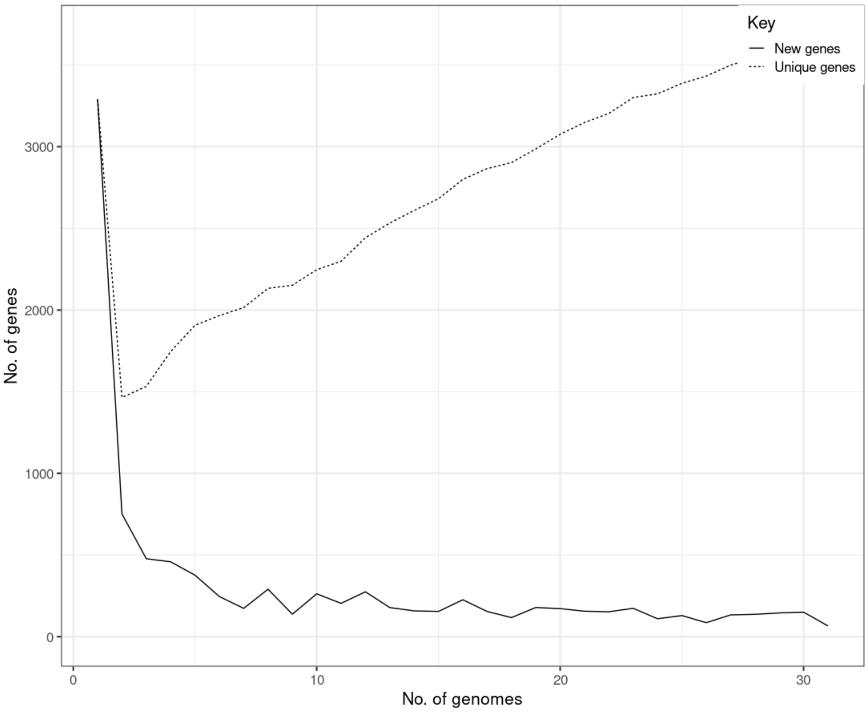


Figure 3.6: Unique vs New genes

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] Ivica Letunic, Peer Bork, *Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation*, Nucleic Acids Research, Volume 49, Issue W1, 2 July 2021, Pages W293–W296, https://doi.org/10.1093/nar/gkab301

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

[14] Treangen, T.J., Ondov, B.D., Koren, S. et al. *The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes.* Genome Biol 15, 524 (2014).
https://doi.org/10.1186/s13059-014-0524-x

[15] 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

[16] Price MN, Dehal PS, Arkin AP, *FastTree 2–approximately maximum-likelihood trees for large alignments.* PLoS One. 2010, 5
doi: e9490-10.1371/journal.pone.0009490.

[17] 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

[18] *https://github.com/biobakery/phylophlan/wiki*

[19] *https://help.ezbiocloud.net/gene-frequency-plot-in-pan-genome/*

[20] Berger FK, Schwab N, Glanemann M, Bohle RM, Gärtner B, Groesdonk HV. *Flavonifractor (Eubacterium) plautii bloodstream infection following acute cholecystitis.* IDCases. 2018 Oct 28;14:e00461.
doi: 10.1016/j.idcr.2018.e00461.