Analysis on SGB4964

Computational Microbial Genomics

Mirolo Filippo, Ossanna Vittoria, Pilli Alessandro

October 20, 2023

1 Introduction

Microbiomes are present in and on the human body: based on the body district we consider, we could find deep differences linked to that certain environment. Among them, the most studied human microbiome is the one living in the gut: this has been widely investigated for its relevance for human health. Indeed, the majority of the samples currently available come from the digestive system, also, it is one of the most populated. It is often difficult to grow bacteria coming from the human microbiome in the lab: many anaerobic bacteria that die in the presence of oxygen, many need complex media or other organisms to survive. To this day, we still do not know whether it is impossible to grow them or - most probably - we just have not yet found the ideal conditions for this to happen (uSGBs). Therefore, we need to rely on computational approaches to study them: the method that is currently most used for metagenomic DNA sequencing is Shotgun. The pipeline of this process consists in a first part which is carried on at a wet-lab level and then a computational analysis. For these reasons, many SGB still are not characterized and still remain a black box after the sequencing, curation and binning processes. Here, we are analyzing SGB4964, a species-level genome bin that consists of 31 metagenomic assembled genomes. Our focus is to characterize this specie with information about protein annotation, pangenomic analysis, phylogenetic trees for both core and accessory genes and genetic profiling.

2 Methods

In order to get annotation of the MAGs corresponding to the SGB4964, we used Prokka: a command line software tool to fully annotate a draft bacterial genome that produces standards-compliant output files for further analysis or viewing in genome browsers [1]. This tool exploits several databases and secondary programmes to perform annotation of the genome that it is given as input. This first analysis has been done by running Prokka with the option `--kingdom Bacteria` on all MAGs present in SGB4964. For each MAG, Prokka generated a General Feature Format file, a summary and a table with the obtained annotations. From the latter two files, we extracted information about the amount of coding proteins, known proteins and hypothetical proteins through a python script. Scripts are available at this GitHub repository.

The next step consists in pangenomic analysis with Roary, a high speed stand alone pan genome pipeline, which takes annotated assemblies in GFF format (produced by Prokka) and calculates the pan genome [2]. In order to run Roary, we set the minimum percentage identity for Blastp

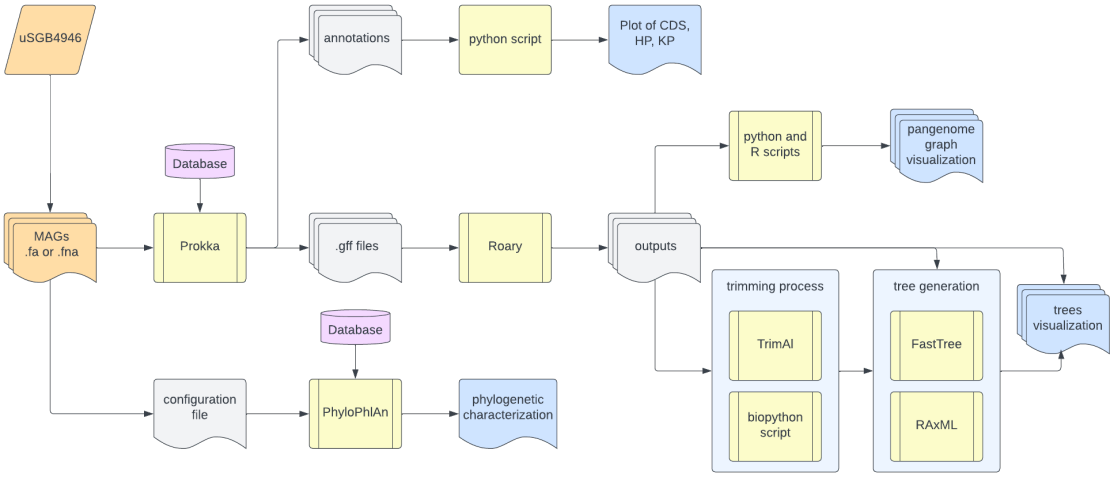


Figure 1: Overall pipeline of the project

to `95`, while the threshold for defining a core gene is set to `90` (percentage of isolates a gene must be present in). In addition we used the options for fast core gene alignment with MAFFT and for creating a multiFASTA alignment of core genes using PRANK (`-e -n`). From the files that this tool returns we get graphical representation of core and accessory genomes, as well as pangenomic analysis. These plots are obtained through R and python scripts linked to the Roary web page (the R script maintained with the standard parameters and the python one was adjusted to match ours). Roary will also produce a core gene alignment, in our case is used to get a phylogenetic tree for accessory genes and (after further downstream analysis) also phylogenetic trees for core genes. In order to produce the final tree from the core gene alignment, we analyze several outcomes generated through combinations of different processing of not trimming or trimming (with TrimAl [3] or a biopython script that analyzes the columns that are not equal available on the GitHub of this project) and tools of tree generation. To obtain the tree we experiment both FastTree - which infers approximately-maximum-likelihood phylogenetic trees from alignments of nucleotide or protein sequences [4] - and RAxML, another popular program for phylogenetic analysis of large datasets under maximum likelihood [5]. For the trimming with TrimAl fraction of sequences with a gap allowed to `0`, while the minimum average similarity allowed is set to `1`. In order to use FastTree we used the command with the nucleotide option (`-nt`), while for RAxML we used only standard optimization for processes (`T 4 -m GTRCAT -p 42`). The graphical representation of the trees has been created through FigTree [6] or iTOL [7]

From the initial MAGs, we also got a profile with PhyloPhlAn, an integrated pipeline for large-scale phylogenetic profiling of genomes and metagenomes [8]. The setup for this tool includes `Diamond` as the main amino acids database and DNA mapping method, `Mafft` for multiple sequence alignment, the usage of `TrimAl` and the construction of a tree using `RAxML`.

A schematic representation of the pipeline is shown in Figure 1.

3 Results and Discussion

3.1 Description of the bin

There are 31 MAGs assigned to the SGB4946: all of them come from stool samples of westernized individuals. 27 out of 31 come from healthy individuals, aside from them, the ill samples come from subjects with adenomas, type 2 diabetes and an asymptomatic subjects (no reference disease is in this last case reported). Around a half of samples come from female individuals, a fourth from male while for the rest we have no annotation available.
The majority of the samples have been obtained through IlluminaHiSeq sequencing, while for three of them we do not have any annotation. Individuals taken into account in this study come from different countries: one from China, three from Austria, two from Sweden, seven from Denmark, one from Spain, one from Mongolia, one from USA and 15 from the United Kingdom.

Each MAG present is a high quality bin, therefore they have a completeness above 90% and a redundancy below 5%.

3.2 Genome annotation

After running Prokka on every MAG present in our SGB, we analyzed the annotation through a python script. The graph in Figure 2, reports for each MAG the total number of coding sequences found in the annotations (CDS), the number of known proteins (KP) and the number of hypothetical proteins (HP). For each sample, we get a number of KP always higher than the HP, the respective averages are 1120 for KP and 932 for HP. The amount of CDS, hypothetical proteins and known proteins seems to vary accordingly: it is clear that the amount of hypothetical proteins and known proteins annotated depends on the overall CDS detected. Nevertheless, the number of CDS annotations seems not to correlate with the amount of reads from the sequencing nor the number of bases. This is consistent with the definition and creation of high quality MAGs: the number of reads should not be a discriminant once we get to the definition of an high quality metagenomic assembled genome.

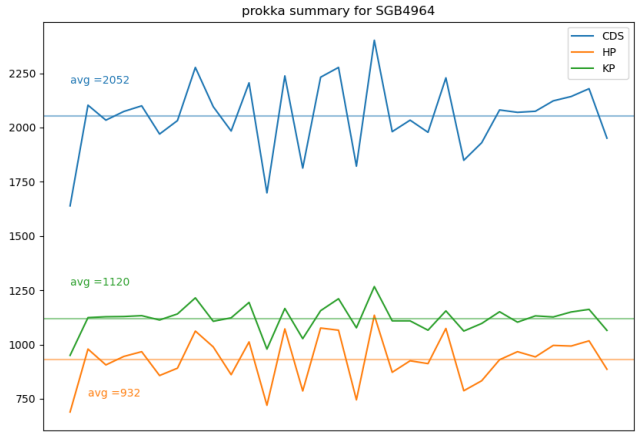
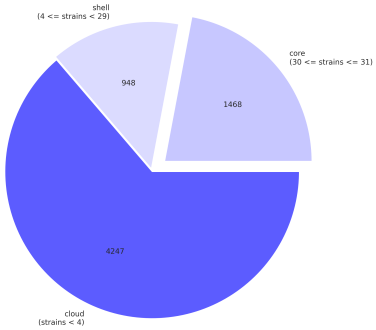


Figure 2: Prokka summary analysis for SGB4964. In blue the trend of coding sequences, in green known proteins and in oragne the hypothetical proteins annotated with the tool.

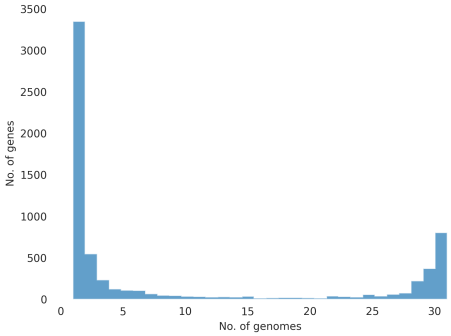
3.3 Pangeome Analysis

The pan genome analysis performed with Roary as described in the methods section gives information about the overall composition of the genome. A gene is selected as a core gene if it is present in between 30 and 31 genomes, to the shell genome if in 4-29 genomes, or to the cloud genome if present in 0-3 genomes. As well represented in the pie chart and the corresponding histogram (Figure 3).

From Figures 4 obtained from the R script, we can observe that an increase in the number of considered genomes corresponds to an increase in the number of total genes and a stabilization of the conserved genes that represent the core gene and reflects the number obtained in the pie chart. On the contrary, as the unique genes keep increasing the number of new genes decreases to a range of 10 to 30 genes. From the first graph we can infer that the pan genome is open, since the number of total genes keeps increasing and the accessory genome heavily outweighs the core genome, but we can not entirely exclude a closed pan genome, due to the decreasing number of new genes from the second graph and the relatively low number of samples available.

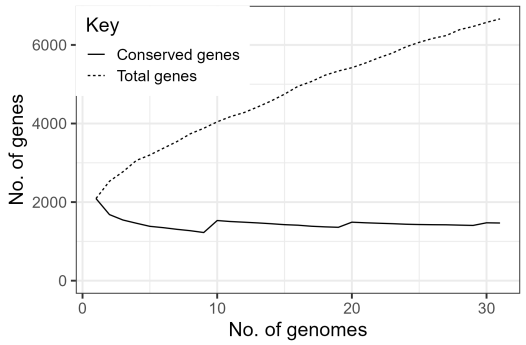


(a) Pie-chart illustrating the pan genome. The cloud genome is composed of 4247 genes; the shell genome contains 948 genes, and the core genome 799.

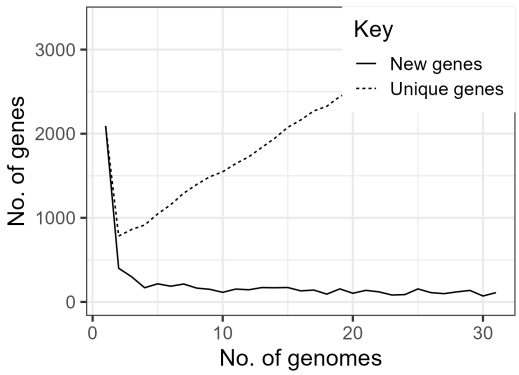


(b) Histogram reporting the number of genes included in different numbers of genomes. It can be observed that around 1500 are shared by 30 or more genomes (the core genome) while a larger amount of genes are shared by fewer genomes.

Figure 3: Roary pan genome analysis from the python script



(a) Number of total genes (pan genome) and conserved genes (core genome) with respect to the number of genomes.



(b) Number of new genes (accessory genome) with respect to the number of genomes.

Figure 4: Roary pan genome analysis from the R script

3.4 Phylogenetic Analysis

From Roary we obtained an accessory genes tree which is based on presence and absence of items and a core gene alignment. From the latter, we used RAxML in order to get the final tree. This alignment has been also trimmed with a biopython script and again used in RAxML. All graphical tree generation reported have been generated from FigTree (Figure 5).
The phylogenetic analysis is based on the comparison of the trees, two of them are given by Roary. One is constructed in function of the presence or absence of the accessory genes, the other one with the differences on the core genome alignments (Figure 6). A third tree was done after the trimming of the data from Roary, with the use of a Python script and RAxML (Figure 7). Additional information to understand the phylogeny of our MAGs have been retrieved thanks to the use of PhyloPhlAn.
From the first tree we can already see some clustering, in fact two main branches are present. Confronting this tree with the one from the alignment of the core genes it is possible to underline that the main structure is conserved even though there are changes in the smaller branches. From these two trees the one constructed from the core genes is more reliable because due to the distances and the phylogeny is measured on the same genes. The last tree displayed has a similar pattern to the core gene tree, but once again we can find many differences as a consequence of the trimming performed by Python.
PhyloPhlAn in this case search was used against a reduced database to look for the microbe with the highest similarity with our MAGs. As concerns our sample, the analysis return a microbe of the Lachnospiraceae Family as the most similar. The distances of all our MAGs are analogous, which means that all our genomes are similar. In conclusion the distances reported by PhyloPhlAn are coherent with the trees constructed.
With the use of the supplementary data of the samples we tried to find some common features that can group the MAGs in a comparable way with the clusters made by the trees, but we did not find any relevant conserved pattern.

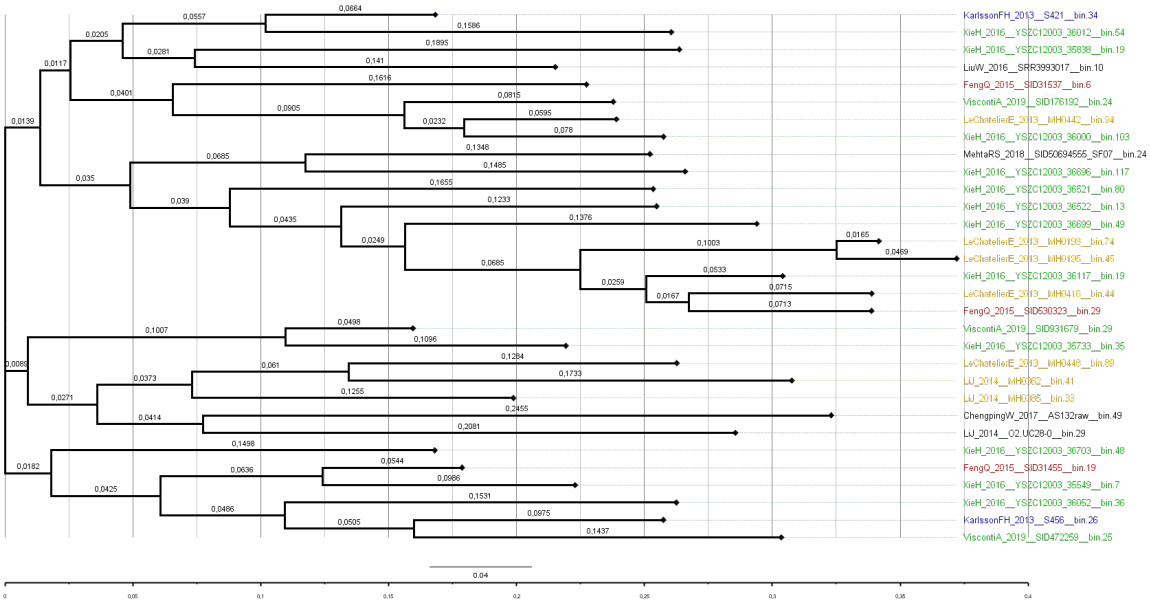


Figure 5: Tree generated from Roary through presence and absence of accessory genes. The colors of the samples represent the provenience of the sample: no clustering based of this labels appear from this tree nor from the next ones. In green we represent UK samples, red for Austria, yellow for Denmark, blue for Sweden, while in black we represent samples with a unique provenience. The same labelling method is applied on the following trees.

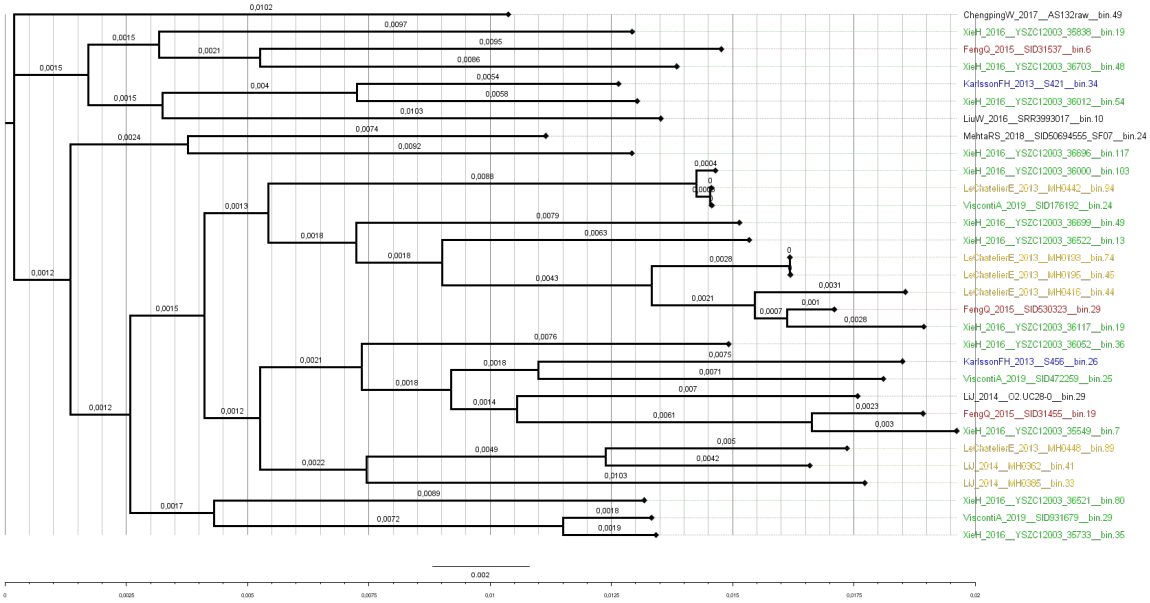


Figure 6: Tree generated from Roary through core genes. The colors of the samples represent the provenience of the sample.

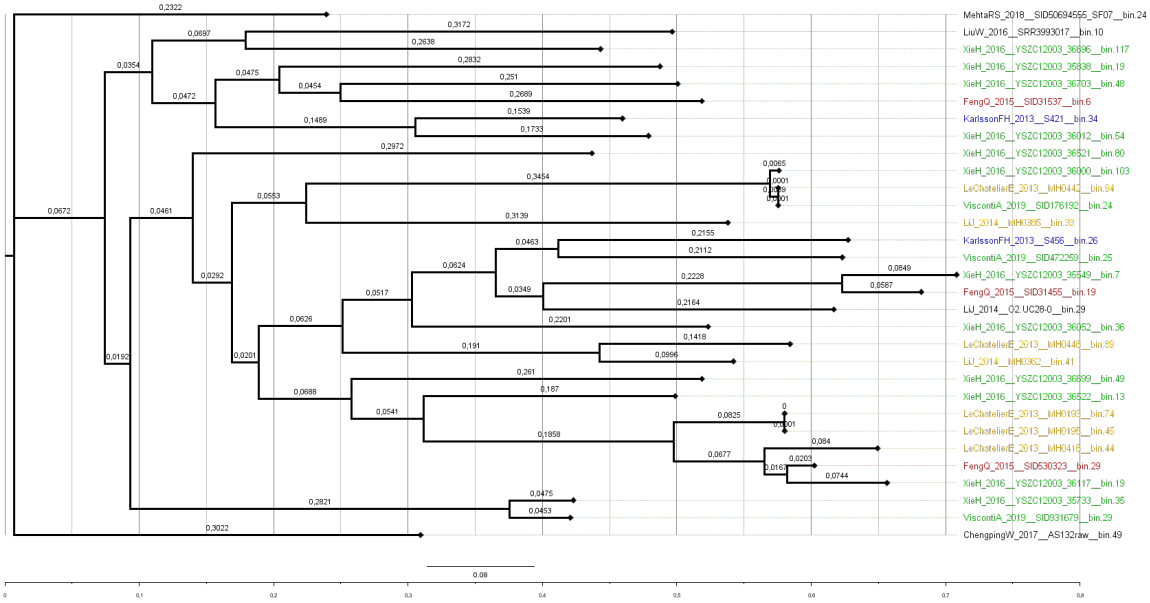


Figure 7: Tree generated from Roary through core genes after trimming. The colors of the samples represent the provenience of the sample.

4 Conclusions

In conclusion the metagenomic analysis of this SGB established an association with the phylogenetic family of Lachnospiraceae, a common gut associated microbe. In literature it is generally linked with an inflammatory response and this can be a starting point for further analysis on the identification of this possible new species. Also, from our data it seems that SGB4964 has an open pan genome, while the numbers of known proteins annotated are always higher than the hypothetical ones.

References

[1] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 03 2014.

[2] Andrew J. Page, Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T.G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693, 07 2015.

[3] Salvador Capella-Gutiérrez, José M. Silla-Martínez, and Toni Gabaldón. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973, 06 2009.

[4] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution*, 26(7):1641–1650, 04 2009.

[5] Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 01 2014.

[6] Kristof Theys, Philippe Lemey, Anne-Mieke Vandamme, and Guy Baele. Advances in visualization tools for phylogenomic and phylodynamic studies of viral diseases. *Frontiers in Public Health*, 7:208, 08 2019.

[7] Ivica Letunic and Peer Bork. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research*, 47(W1):W256–W259, 04 2019.

[8] Francesco Asnicar, Andrew Maltez Thomas, Francesco Beghini, Claudia Mengoni, Serena Manara, Paolo Manghi, Qiyun Zhu, Mattia Bolzan, Fabio Cumbo, Uyen May, and et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using phylophlan 3.0. *Nature News*, May 2020.