Characterization of a uSGB: SGB4936

Randazzo Giulio, Leufen Jan, Zardo Marco
University of Trento
giulio.randazzo@studenti.unitn.it, jan.leufen@studenti.unitn.it, marco.zardo@studenti.unitn.it

***Abstract* – The goal of this paper is to do a characterization of the uSGB4936 which contains bins from four different studies which collected stool samples between individuals from Great Britain and the Fiji Islands. The samples from Great Britain were labeled as westernized, whereas the samples from the Fiji Islands were labeled as non-westernized. Main parts of our analysis included a taxonomic analysis, analysis of core-genome size dependent of westernization and phylogenetic analysis.**

I. INTRODUCTION

Computational and biotechnological improvements in the field of metagenomics, whose high-throughput techniques cover a remarkable part, have increased the level of the analysis of microbial communities and the knowledge of new microbial species (taxa) and genes involved in different environments such as human microbiome.

Although most classification of these ecosystems rely on culture-independent genomic procedures to try to expand the expertise of an increased set of microbes in different settings, this technique does not take into account the huge fraction of microbial diversity which is still unknown/unexplored. One of the advancements of the Metagenomic evolution is the possibility to overcome this gap between characterized-uncharacterized differentiations **by studying genetic material sampled directly from the environment**. As starting point, the high-throughput sequencing marks the total genomic information of a sample to provide a huge amount of fragments in order to catch the entire complexity of the environment showing in the sample.

The shotgun sequencing is followed by Assembly, where the original sequence is reconstructed from the short reads previously obtained, thanks to the overlaps between them. As a result, the longer composite sequences known as contigs are assembled without gaps, with a high-quality DNA stretches spanning > 1K. The third step is the collection of all the contigs into MAGs (Metagenome-assembled Genome). The generation of the MAGs is obtained thanks to the binning techniques, that represent the "best effort" to identify the consensus sequences within a certain taxonomic group.

This procedure usually requires large computing power as well as an outstanding runtime for many factors. First, the fact that the input content needed by the binning approach comes from a complex environmental sample, with mixed communities of microbes such as fungi, viruses bacteria and archea. Where in most of the case, completely sequenced referenced genomes of closely related taxa are not available; even if they are required for a taxonomic assignment based on sequence similarity. Moreover the time spent during such computations increases proportionally to the number of contigs previously obtained and to the absence or presences of some reference sequence collection, in which it has to count also the size of them.

Thus, to make the taxonomic assignment possible the input data has to be linked to their MAGs with clustering algorithms. However the selection between clusters or classify based-method is strictly correlated on the availability of some reference data. In presence of them, classification became more precise; on the other side, clustering allows detection of taxonomic bins which otherwise would go unobserved [2].

Binning procedure can be produced manually thanks to particular software as Anvi'o that obtains more accurate results but loses throughput and speed, or automatically, using software such as MetaBAT.

The main aspects that are used for grouping MAGs are the Sequence composition analysis, which joins comparison of K-mers frequencies among contigs and clustering methods to retrieve information about similarity patterns among contigs and differential coverage, using to compare the relative coverage of contigs in order to assign them to the same MAG if their coverage is similar within the sample and covaries among the whole set of samples.

The produced assembled genomes have been checked out to define their quality, taking into account the bacterial single-copy genes completeness and redundancy. In the final output, the bins are grouped together in the same species genome bin (SGB) if they have followed a certain threshold of nucleotide identity.

II. BIN'S DESCRIPTION

The collection of high-quality MAGs managed for this project refers to the uSGB labeled SGB4936. Investigating the statistics associated with the bins, showed on the bin data TSV files, we discovered 32 MAGs with an average completeness of 95.98% and redundancy of 0.58%. These values approve the high-level of the data under analysis, since the thresholds of standard drafts which define a completeness > 90% and a redundancy < 5% are well denoted [1]. On the other side, we examined the metadata relating to the origin and the gender of the samples before starting genomic analyses in order to detect some useful information to better describe the analysis we are going to perform. As a result, we found that the species we are taking into account could be defined as a gut colonizer, since all the reconstructed genomes come from stool samples: almost all of these are hailed from Europe, in particular Great Britain, but also from non-westernized countries [5]. Moreover the majority of samples were being taken from female cohort where a small fraction of these were provided by more than 1000 volunteers from twinsUK, the largest twin registry in the world [13]. The study conditions were mainly characterized by healthy people in different ages: meaning that the uSGB

represented a putative specie well distributed in the human gut in different stages of human's life. Last but not least, Illumina Hiseq was the sequencing platform used for all the files where the NGS technology was indicated, with a median read length across the samples of 100 and a small fraction of them of 125.

III. METHODS

A. Genome Annotation with Prokka

The first step of analyzing the genome data we were given is to annotate it with the tool prokka [11]. Prokka can be used to annotate bacterial genomes as well as archaeal and viral genomes. We can therefore use it to annotate our suspected bacteria genomes/MAGs from the uSGB called SGB4936. Prokka will identify all coding sequences in our genome and compare it to its database, which contains a list of already known proteins. Unknown but suspected proteins will be labeled as "hypothetical proteins". Prokka will also provide us with metadata information, which includes among other things the amount of bases and contigs. Finally we can use the output of prokka to analyze the amount and ratio of known to hypothetical proteins.

The command for annotating our genomes was encased in a loop so that it was carried out for all our given MAGs. Prokka uses FASTA files as input that contain all the contigs of one MAG in one file. The command is shown in figure 1. We specify the kingdom that we expect, which is Bacteria in

```
for mag in `ls | grep f.*a`
do
    prokka --kingdom Bacteria --outdir ${mag}_out
     --prefix ${mag} ${mag}
done
```

Fig. 1. Prokka command used for annotating the MAGs of our uSGB. The commands runs in a loop over all the FASTA files in the directory.

the case of our MAGs. As output of this command we get folders that contain the annotated genomes as well as metadata information. This includes a table of all found proteins with their names or the label "hypothetical protein" if it was not known to prokka.

B. Pan-genome Analysis with Roary

After all our MAGs are annotated, we can analyze the complete pan-genome that consists of all the genes of all MAGs combined. For this analysis, we use the tool roary [9]. Roary will look at the annotated genomes for each of the MAGs one after the other. For each genome, it will look for genes that it did not see before (new genes) and for the already seen genes it will keep track of the amount of genomes this gene was contained in. Genes that can be seen in more then 90% of genomes we will consider as core genes. We choose a number lower then 100% because it can always happen, that for one of our MAGs the genome has sequencing errors or missing pieces which result in core genes missing. The more genomes we add to our analysis, the more we run the risk of discarding core genes because of faulty data. The roary command we used can be seen in figure 2. The parameter `*.gff` is a regular expression over all input files with file ending `gff`. The second parameter `-f` determines the output directory. The parameter

```
roary *.gff -f roary_output_en -e -n -i 95 -cd 90
     -p 10
```

Fig. 2. Roary command used for creating pangenome and multiple core-gene alignment

`-e` will trigger roary to do a multiFASTA alignment of the core genes using the PRANK [8] algorithm. This algorithm will align the core-genes and will take the codons into account. It is very accurate, but also very slow. Therefore we also add the `-n` parameter that will use mafft [6] instead of PRANK for the core-gene alignment. Mafft is based on nucleotides and as accurate as PRANK, however it is much faster. Parameter `-i 95` sets the percentage idendenty for blastp. Above these percentage two sequences are viewed as identicaly. This value may be reduced if species are very diverse. The parameter `-cd 90` sets the above mentioned threshold for core-gens to 90%. The last parameter `p` sets how many threads roary uses to run the command [12].

So far we have analyzed only the complete set of MAGs. We were also interested in the change of the core-genome when removing a certain part of the MAGs from the analysis. This part was the 7 MAGs that came from non-westernized individuals. The reason for this is explained in the discussion section, and here only a technical explanation is given. In order to do the analysis, we removed all seven `.gff` files from the non-westernized individuals and rerun the roary analysis. To get a base value of how many more core-genes we can expect in general when removing seven of or MAGs we rerun the test further ten times while removing seven different random MAGs every time.

C. Phylogenetic Structure with Roary and FastTree

The default roary output contains the phylogenetic tree that was generated based on the accessory genes of the pangenome. This file can be displayed directly with tools like iTOL [7]. When running roary with the added parameter `-e` as we did above, we also get the alignment file that contains the alignment of the core-genes. Based on this alignment, we can use the software FastTree [10] in order to generate another phylogenetic tree, which is then based on those core-genes. We used the command shown in figure 3, where `-nt` stands for nucleotides, as we are working on nucleotides.

```
FastTree -nt < core_gene_alignment.aln >
    core_gene.tre
```

Fig. 3. FastTree command used for generating phylogenetic tree

D. Taxonomic Assignment with PhyloPhlAn

The last step in our analysis was the taxonomic assignment of our uSGB. The tool we used is PhyloPhlAn [3]. The database that we used for the taxonomic assignment was given through the project. The command we used can be seen in figure 4 The input for PhyloPhlan is given as the parameter `-i input`, whereby `input` is a folder which contains all FASTA files that should be taxonomically analyzed. The parameters `-nproc 4 -verbose -o phylophlan_output -database_update` are all of technical nature and define the number of threads, set the verbosity of the output, configure

```
phylophlan_metagenomic -i input -o
    phylophlan_output --nproc 4 -n 1 --
    database_update -d CMG2324 --verbose
```

Fig. 4. PhyloPhlAn command used for getting the taxonomic assignment of our uSGB

the handling of the database and set the output directory respectively. The parameter `-n 1` how many of top results should be outputted. By passing `1` we only get the top result. Finally the `-d CMG2324` parameters gives PhyloPhlan the name of the database to download.

IV. RESULTS and DISCUSSION

A. Genome Annotation

Prokka took as input the FASTA files and it generated a text file containing the number of Coding DNA Sequences (CDS) and contigs of each sample, plus other annotation details used to retrieve the number of hypothetical and known proteins. The data is shown in table I. From table I, it is possible to observe that the number of annotated proteins remains approximately constant over the samples (mean: 1488 ± 60; median: 1505), and the same happens to the number of hypothetical proteins (mean: 1306 ± 95; median: 1308). The mean and the median values of the known over the total proteins and of the hypothetical over the total proteins were respectively 0.5331 ± 0.012 and 0.5330, 0.4669 ± 0.012 and 0.4670.

Statistics/Parameters	**CDS**	**hypo**
Mean ± s.d.	2794 ± 146	1306 ± 95
Median	2809	1308
Statistics/Parameters	**CDS**	**hypo**
Mean ± s.d.	1488 ± 60	0.4669 ± 0.012
Median	1505	0.4670

TABLE I

Table 1. Basic statistics of our MAGs

B. Pangenome Analysis

The Pangenome analysis provided us the overall composition of the putative genome in terms of estimations of core genes and accessory genes. The allocation of a gene to core genome happens if that gene is shared in a particular range of the genomes taking into account: for instance, gene belonging to core genome is shared to 31 or 32 genomes (almost the whole present in our SGB), while genes coming from soft core genome and shell are distributed in 30 genomes and in 4 – 29 genomes respectively. On the other side the genes linked with the cloud are more characterized in less genomes (0 up to 3) since they usually might encode for some special functions which do not have a high distribution within all the other genomes in the sample. The visualization of this data can be achieved using a pie chart and a histogram thanks to the python script present in the Roary Github repository [9]. In the figure 5, we can see that the majority of the genes has been assigned to the accessory genome (around 3500 genes), however the core genome represents more than 20% of all the defined genes and together with the shell and the soft core one accounts for about 48% of the overall genes in the considered MAGs. This last statement underlines a significant balance between the variability showed by accessory genes and the number of shared ones, strictly connected with a conservation of some characteristics at the species-level. A similar trend can be also recognized in the figure 6, where we report the amount of genes figured out considering a given number of genomes.

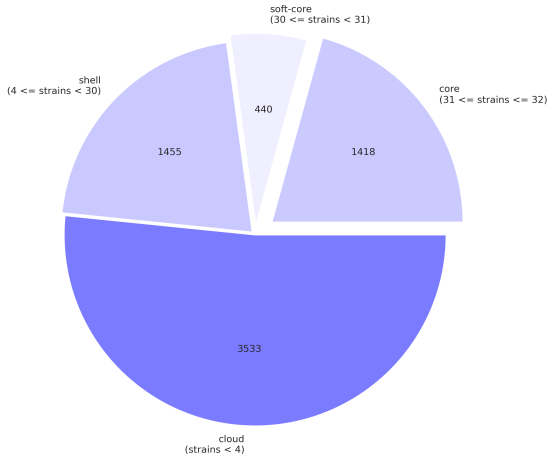

Fig. 5. Pie-chart illustrating the pangenome composition obtained from Roary. The cloud genome is composed of 3533 genes; the shell genome contains 1455 genes, the soft-core genome 440 and the core genome 1418.

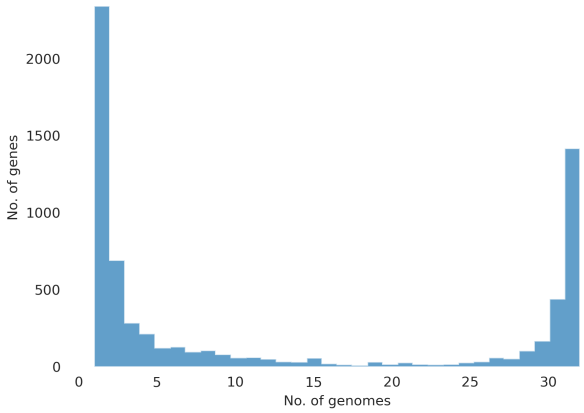

Fig. 6. Histogram reporting the number of genes included in different numbers of genomes. It can be observed that around 1500 are shared by 30 or more genomes (the core genome) while a larger amount of genes are shared by fewer genomes.

As we can see, there are two spikes in the plot. The first shows up the strength of cloud genome: a larger number of genes shared by few genomes. The second emphasizes the block of the conserved one: around 1500 genes are in common with more than 30 genomes. The representation of the core-genome and the pan-genome size at an increasing number of genomes was consistent for us to demonstrate that

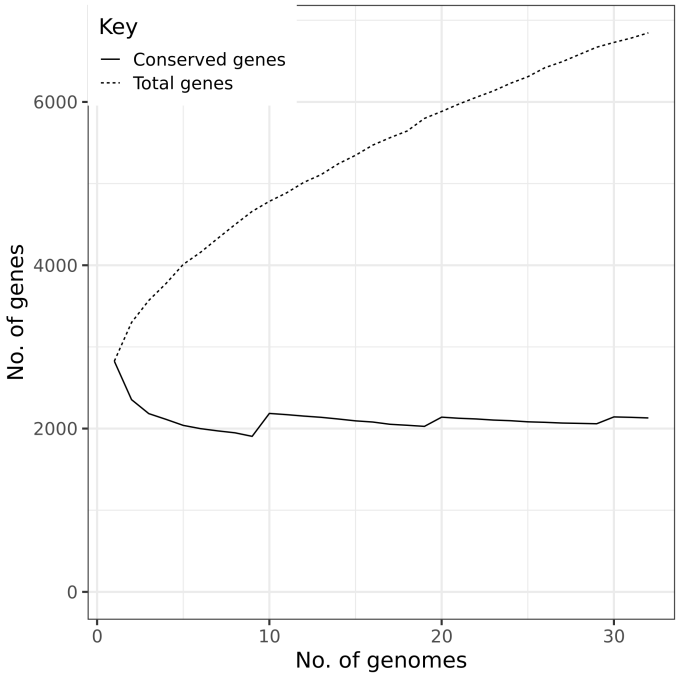

Fig. 7. Number of total genes (pangenome) and conserved genes (core genome) with respect to the number of genomes.

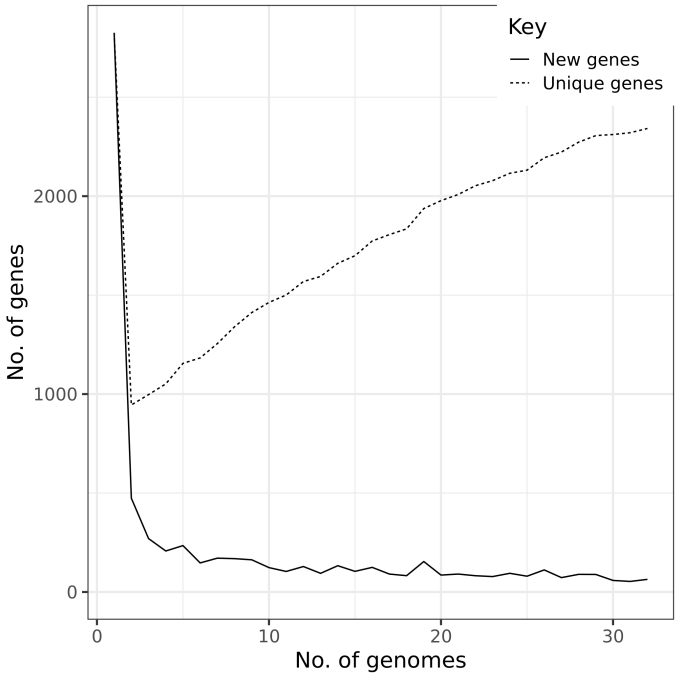

Fig. 8. Number of new genes (accessory genome) with respect to the number of genomes.

our uSGB shows some features firmly connected to a species with an open pan-genome but with some considerations to infer. Figure 7 shows that the number of genes in the accessory pan-genome keeps increasing, whereas the number of genes in the core genome remains stable with a number close to 2000 genes. Figure 8 confirms the same behavior, since the number of unique genes found when considering increasingly more genomes continues to increase. However all the visualizations shows a common characteristic: the function of the total genes in fig 7 and the function of unique genes in fig 8 tent to reach a plateau, meaning that the size of accessory genome might be not boundless, changing the pan-genome definition previously defined. All the plots were obtained using the R script present in the Roary Github repository and they are used to validate our assessments.

C. Core-genome Size dependent on westernization

We saw a hint in the overall core genome plot of roary that there might be a trend in the non-westernized genomes to have less core-genes present than the other genomes from the westernized samples. To further examine this assumption, we decided to do a test set of roary run with different MAG sets as input. We decided on the following test suite:

- Run roary with all input MAGs (32 MAGs)
- Run roary with only the westernized MAGs (25 MAGs)
- Run roary with 10 times with seven random MAGs missing (10x 25 MAGs)

The goal of first test case was to get the benchmark and the goal of the second test case was to see if the amount of core-genes would increase if we omit the non-westernized MAGs. However these two test on their own would not be very meaningful as we would also expect an increase in core-genome size every time we remove a MAG from the analysis. The result would only be meaningful if we could see a bigger increase in core-genome size only for the non-westernized

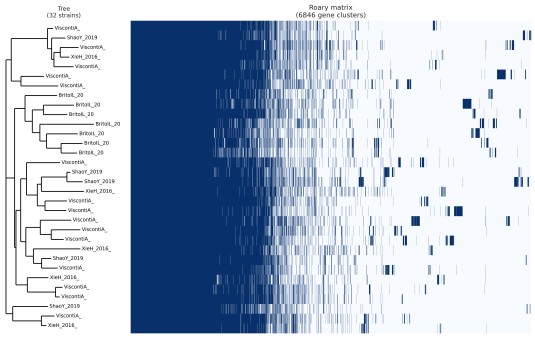

Fig. 9. Figure showing a tree drawn considering the presence/absence of accessory genes on the left, obtained from the Roary matrix on the right.

MAGs removal. The results are shown in figure 10. The baseline of all MAGs included in the analysis is 1418 core-genes. The then runs with seven randomly removed runs have core-genome values between 1423 and 1630 and a median value of 1530 core-genes. As expected, the core-genome size increases when removing MAGs from the analysis. For our ten tests, this increase is approximately $7,9\%$ (maximum 14.9%). The run that only had westernized MAGs included had 1785 core-genes which is an increase of $25,8\%$.

We cannot be sure if this increase is a coincidence or meaningful, but one could speculate that the diet of the non-westernized individuals is different and therefore leads to more diverse individuals on the bacterial levels with less core genes and more specialized genes. It would also be possible that there are a different kind of core-gene set between the MAGs from the non-westernized individuals. This core gene set would not be visible in our analysis, as it may not be shared with the

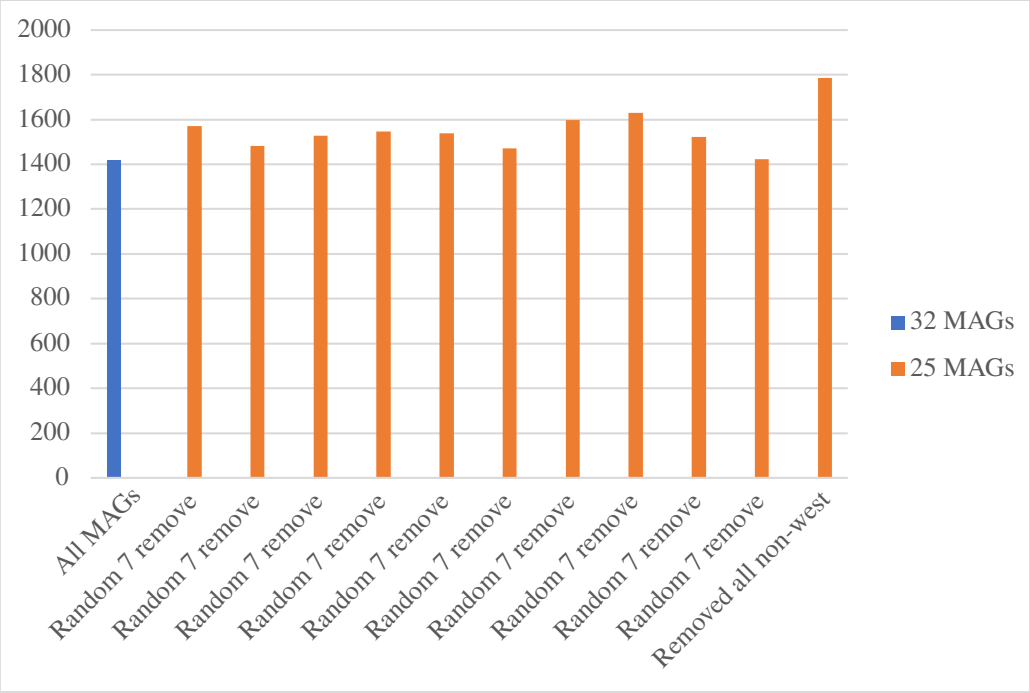

Fig. 10. Core-genome sizes for different sets of MAGs that were analyzed with roary. The first column has baseline value with all MAGs. In the middle we have the ten runs with seven randomly removed MAGs and the final column contains the target value of all non-westernized MAGs removed. The color indicates the amount of MAGs analyzed in each run.

westernized MAGs.

D. Taxonomic Assignment

Taxonomy assignment was performed using PhyloPhlAn. All the MAGs belonging to our uSGB were taxonomically assigned to the *Lachnospiraceae* family, the identified species was *Roseburia hominis*. From the analysis of one of the genomes, we obtained the output seen in figure 11.

```
kSGB_4936:Species:k__Bacteria|p__Firmicutes|
    c__Clostridia|o__Eubacteriales|
    f__Lachnospiraceae|g__Roseburia|
    s__Roseburia_hominis|t__SGB4936
    :0.019991196551943913
```

Fig. 11. Taxonomic result that was outputted by PhyloPhlAn

E. Phylogenetic Structure and Association With Host Data

As mentioned, phylogenetic trees were generated using Roary and FastTree while the visualization was performed thanks to iTOL, an online tool for the display, annotation and management of phylogenetic and others tree directly on the browser. The results turn out two different structures showed in the two figures below. The Figure 12 represents the Tree of the core genes, therefore it was outlined by viewing the alignment of core genes comprised on the .aln file generated by Roary. The output file was reached once it was used fastTree. On the other side, figure 13 was inferred from the presence/absence of accessory genes which it was already present as an output file of Roary. Then the representation of the phylogenetic trees was enriched with the addiction of colors (red block) around the names of the samples to distinguish the non-westernized and the westernized ones. All the efforts were performed in order to associate phylogeny with host data, and we observed some similar features among the trees.

V. CONCLUSION

Thanks to this project, we have appreciated how the analysis of the sequencing data coming from an unexplored bacterium works. The first assessment we performed was the evaluation of the quality of the given uSGB and the computing of some statistics on the associated metadata. after that we gained some information about the annotated MAGs, in particular the number of coding sequences and the proportion of hypothetical proteins in these coding regions (table I). Considering the classification of the pan-genome, we discovered it to be open, since the amount of total genes keeps to increase if the number of genomes increases as well, while the number of core genes decreases a bit; however the function of total genes tents to reach a plateau in the last part of the plots meaning that it might be necessary including more sequencing data on the already present in order to see if the trends will be validated. Regarding the amount of accessory genome and the core genome, we observed that around the 52% of the pan-genome size is covered by the cloud, but an important fraction is showed as a conserved region, in fact grater the 20% of genes is shared by 90% or more of our MAGs. These results are in line with the taxonomic assignment of the uSGB which it was associated to the *Lachnospiraceae* family: the microbes of this family shows a great inter- and intra-species diversity which increases their impact in the human healthy hosts as an abundant member of the microbiota [4]. Afterward, we examined the phylogenetical features of the MAGs to try to extract some pattern in association with the host data. The huge amount of our samples come from European subject, in particular from

Great Britain, while a fraction of seven is originated from Asia. Interestingly, the identification of the clusters based on the geographical information was in general clear. We were able to identify evolutionary similarities regarding the non-westernized samples: in both the phylogenies, one based on core genes and the other on accessory genes, the cluster connected to the non-European samples is indicated and well-defined with respect to the other groups of MAGs. Moreover the associations between genomes were retained confirming a sort of evolutionary connections shared by the Mags in terms of conserved genes and the new acquired ones. For instance, the sample named XieH YSZC12003 (bin 31) is close to the ViscontiA SID542607 (bin 31) in each of the phylogeny and all from UK. ShaoY samples SID64696f4e and SID64053380 differ a little only in terms of accessory genes and the same is true for the ViscontiA SID498259 and ViscontiA SID365684 samples. A more systematic sampling, with connection with a more detailed report for the host metadata would be required to make some stronger claims.

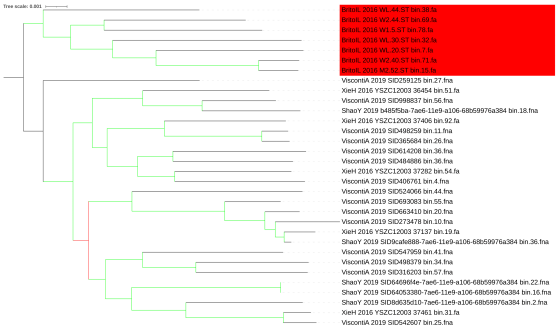

Fig. 12. Tree constructed taking into consideration core genes: the red block showing the non-westernized cluster. The colors in the tree symbolize the certainty with which roary made the split in the tree.

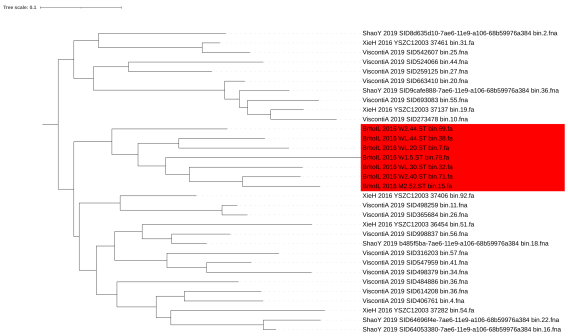

Fig. 13. Tree constructed taking into consideration accessory genes: the red block showing the non-westernized cluster.

REFERENCES

[1] Robert M et al. Bowers. Minimum information about a single amplified genome (misag) and a metagenome-assembled genome (mimag) of bacteria and archaea. *Nature biotechnology*, 35(8), 2017.

[2] McHardy Droge. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Brief Bioinform*, 13(6):646–55, 04 2012.

[3] Segata et al. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communication*, pages 2068–2069, 2013.

[4] Sobara et al. Functional and genomic variation between human-derived isolates of lachnospiraceae reveals inter- and intra-species diversity. *Cell Host and Microbe*, 28(1), 2020.

[5] Brito IL, Yilmaz S, Xu L Huang K, Jupiter SD, Jenkins AP, Naisilisili W, Tamminen M, Smillie CS, Wortman JR, Birren BW, Xavier RJ, Blainey PC, Singh AK, Gevers D, and Alm EJ. Mobile genes in the human microbiome are structured from global to individual scales. *Nature*, 535(7612):435–439, 2016.

[6] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, 07 2002.

[7] Ivica Letunic and Peer Bork. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, 49(W1):W293–W296, 04 2021.

[8] Ari Löytynoja. Phylogeny-aware alignment with prank. *Multiple sequence alignment methods*, pages 155–170, 2014.

[9] 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.

[10] Arkin AP Price MN, Dehal PS. Fasttree 2 – approximately maximum-likelihood trees for large alignments., 2010.

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

[12] Ole Tange. Gnu parallel 20230122 ('bolsonaristas'), January 2023. GNU Parallel is a general parallelizer to run multiple serial command line programs in parallel without changing them.

[13] Rosa F Visconti A., Le Roy C.I. terplay between the human gut microbiome and host metabolism. *Nat Commun*, 10(4505), 2019.