UNIVERSITÀ
DI TRENTO

Analysis and characterization of a uSGB: SGB4894

Elisa Bugani, Stefano Cretti, Maurizio Gilioli

INTRODUCTION

The development of metagenomic tools and pipelines, made possible by the advent of *high-throughput* sequencing techniques, is establishing new genome-wide approaches for the analysis of microbial communities with a relevant impact in many fields, such as epidemiology and ecology. One of the most relevant novelties introduced by this broad field is the possibility to assemble shotgun sequencing data into putative microbial genomes, the so-called Metagenome-Assembled Genomes (MAGs), offering the opportunity to better describe poorly-characterized or never cultivated bacteria. As part of this characterization workflow, MAGs are then assembled into Species-level Genome Bins (SGBs), which are sets of genomes grouped by genetic similarity ($< 5\%$ Mash distance). Among those, unknown Species-level Genome Bins (**uSGBs**) are bins for which no reference isolates and or metagenomically assembled genomes were found in available public repositories. Thus, they represent new putative species which can have a crucial role in expanding microbial diversity and mappability, especially of underrepresented phyla and under-sampled categories and populations. In the context of this study, we analyzed a single uSGB, **SGB4894**, composed of 32 high quality MAGs assembled in the context of an extensive metagenomic study on human microbiome conducted by Pasolli and collegues in 2019[1]. The MAGs contained in this uSGB derive from stool samples of individuals with different ages, gender and health conditions, from different areas of the world, both westernized and non-westernized. The goal of our project was to start characterizing this putative species, by retrieving information relative to genetic features, pangenome and accessory genome, taxonomic annotation and phylogenetic structure; all of this was done by putting into practice a simple metagenomic pipeline, which represents a scalable tool to expand the ability to study the phylogenetic and functional diversity of the microbial world.

METHODS

SGB retrieval

The set of MAGs complete with metadata used for this project refers to the uSGB labeled as SGB4894. As stated before, this uSGB was defined in the context of an extensive metagenomic study[1] and it is available in the Segata Lab website[2]. This uSGB is composed of 32 high quality MAGs, meaning MAGs with completeness $> 90\%$, contamination $< 5\%$ and strain heterogeneity $< 0.5\%$. The SGB was assembled using an all-versus-all genetic distance quantification followed by clustering and identification of genome bins spanning a 5% genetic diversity, consistently with the definition of known species[1].

Genome annotation

Genome annotation on FASTA sequences was performed using Prokka (version 1.14.6). Prokka is a command line software which can be used to fully annotate a draft bacterial genome. Genome annotation is carried out identifying the coordinates of candidate genes and comparing with databases of known sequences, chosen in a hierarchical manner (most trustworthy first). When no match is found the Open Reading Frame (ORF) is labeled as 'hypothetical protein'[3]. To perform the annotation, the following command is used for each MAG:

```
$ prokka --kingdom Bacteria --outdir ${out} --locustag ${tag} \
         --prefix ${mag} ${file_name_of_your_mag}
```

Pangenome analysis

The genome composition of the given uSGB was determined using Roary (version 3.7.0). Roary is a tool that rapidly builds large-scale pangenomes, identifying the core and accessory genes[4]. As input for Roary, the annotated sequences produced by Prokka were used (GFF3 format files). The command used to do the analysis is the following:

```
$ roary *.gff -f roary_out -i 95 -cd 90
```

Only genes present in at least 90% of genomes were considered core genes and a percentage of sequence identity of 95% for BLASTP was used. The output files were then used to obtain the plots shown in the 'Pangenome analysis' chapter of the Results and Discussion part; the graphs were produced using the R and python scripts available at the Sanger Institute Github page[4–6]. Additionally, we used PRANK to produce a multiple sequence alignment of the core genes, which was used afterwards to create a phylogenetic tree[7]. The command for this operation is:

```
$ roary *.gff -f roary_out -e -n -i 95 -cd 90
```

with the flags -e which produces the alignment and -n that allows a fast core gene alignment using MAFFT.

Taxonomic assignment

To obtain a taxonomic assignment of our MAGs we used PhyloPhlAn (version 3.0), a tool that uses species-specific sets of marker genes, identified using UniRef90 gene families, to automatically build accurate phylogenies. Furthermore, the tool assigns a taxonomic label to MAGs based on the NCBI taxonomy, while considering also unnamed and uncharacterized species[8]. In order to do so the following command was used:

```
$ phylophlan_metagenomic -i phylophlan_input -o phylophlan_output --nproc 4 -n 1 \
                         --database_update -d CMG2122 --verbose
```

The command 'phylophlan_metagenomic' was run using as input the folder 'phylophlan_input' containing the genome sequences in FASTA format. The database name was specified using '-d' , whereas '-n 1' specifies the number of SGBs to report in the output.

Phylogenetic analysis and association with host data

Starting from the core genome alignment produced by Roary (the 'core_gene.aln' file), a phylogenetic tree (the 'core_gene_tree.tre' file) was produced using FastTree (Version 2.1.11), a tool for constructing large phylogenies which works by storing sequence profiles of internal nodes in the tree that are then used to produce Neighbor-Joining[9]; to assess the reliability of the trees, FastTree uses local bootstrapping. The following command was used:

```
$ FastTree -nt < core_gene.aln > core_tree.tre
```

A second file produced by Roary ('accessory_binary_genes.fa.newick'), containing information about the presence and absence of accessory genes in each MAG, was used to create a second tree. Trees were visualized and annotated using iTOL (Version 6), an online and freely available tool for phylogenetic tree display[10].

RESULTS AND DISCUSSION

Genome annotation

Running Prokka with the FASTA files as input, we obtained a text file containing the number of Coding DNA Sequences (CDS) and contigs of each sample, plus some other annotation data used to retrieve the number of unknown and known proteins; the results are plotted in figure 1. From table 1 and figure 1, it is possible to observe that the number of annotated proteins remains approximately constant over the samples (mean: 1382 ± 78; median: 1380), while the number of hypothetical proteins seems to fluctuate more (mean: 1359 ± 177, median: 1311). The mean and the median values of the known over the total proteins and of the hypothetical over the total proteins were respectively 0.5058 ± 0.186 and 0.5097, 0.4942 ± 0.186 and 0.4903. The indicated statistics are summarized in table 1:

Statistics/Parameters	CDS	hypo	known	hypo/ratio	known/ratio
Mean	2741 ± 250	1359 ± 177	1382 ± 78	0.4942 ± 0.186	0.5058 ± 0.186
Median	2682	1311	1380	0.4903	0.5097

Table 1. Basic statistics of our MAGs

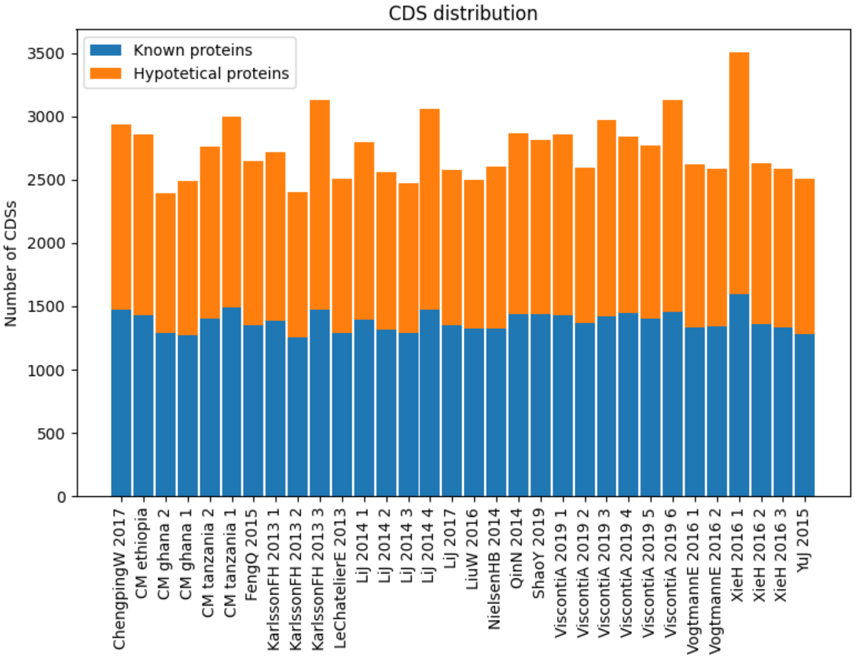
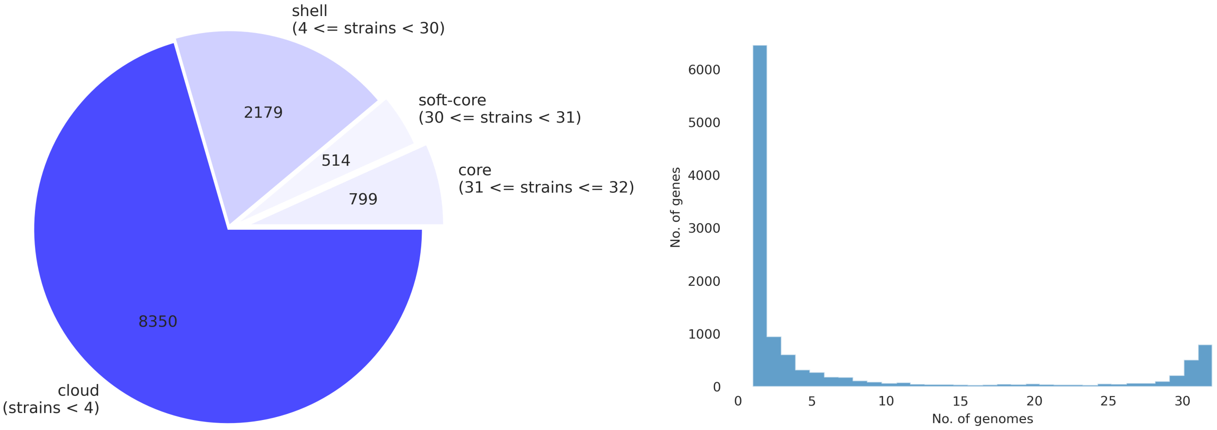


Figure 1. Number of known and hypothetical proteins in each sample

Pangenome analysis

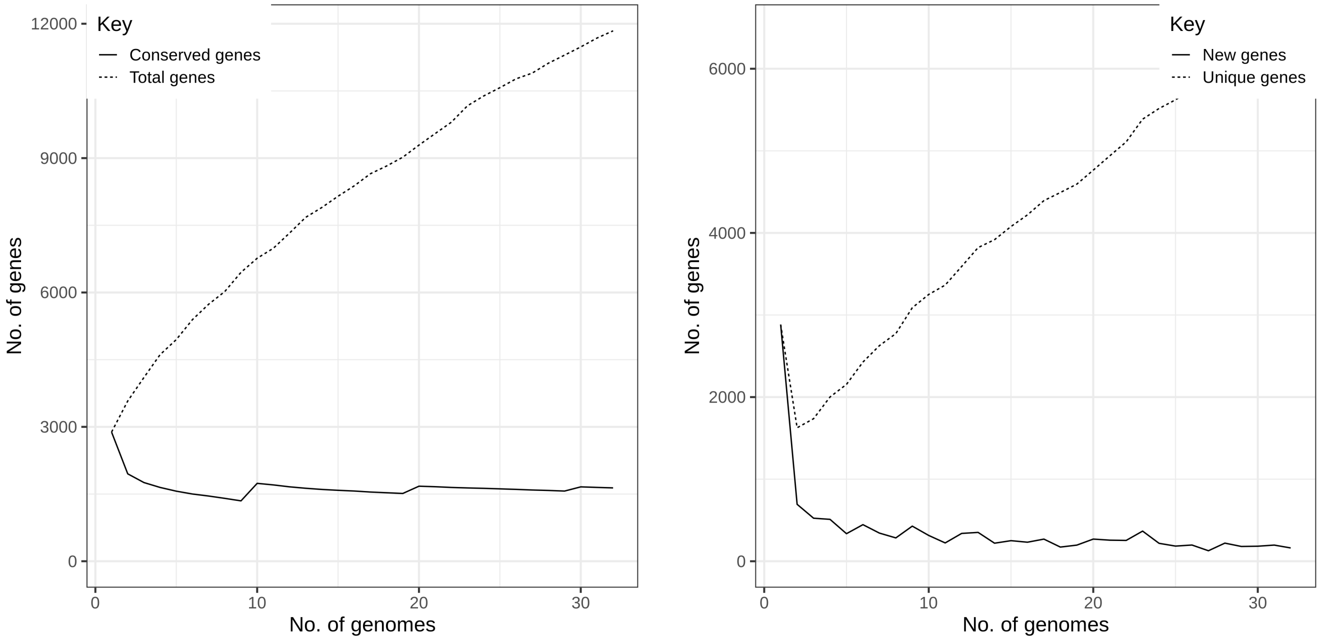
The pangenome analysis, performed using Roary (version 3.7.0) (see the 'Pangenome analysis' section in Methods), generated information about the overall composition of the putative genome. A gene was assigned to the core genome if it was present in 31 or 32 genomes, to the soft-core genome if in 30 genomes, to the shell genome if in 4-29 genomes, or to the cloud genome if present in 0-3 genomes. A representation of this data can be obtained using a pie chart (figure 2a) or a histogram (figure 2b).



(a) Pie-chart illustrating the pangenome composition obtained from Roary. The cloud genome is composed of 8350 genes; the shell genome contains 2174 genes, the soft-core genome 514 and the core genome 799.

(b) Histogram reporting the number of genes included in different numbers of genomes. It can be observed that around 1000 are shared by 30 or more genomes (the core genome) while a larger amount of genes are shared by fewer genomes.

Figure 2. Figures obtained using the python script present in the Roary Github repository[4,6]



(a) Number of total genes (pangenome) 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 3. Plots generated using the R script present in the Roary Github repository[4,5]. An increase in the number of new genes is observed when progressively considering a larger amount of genomes

From figure 3, we see that an increase in the number of considered genomes corresponds to an increase in the number of total genes (3a), as well as an increase in the number of new genes (3b); this indicates that the pangenome is ***open***, since the number of total genes keeps increasing and thus the accessory genome heavily outweighs the core genome.

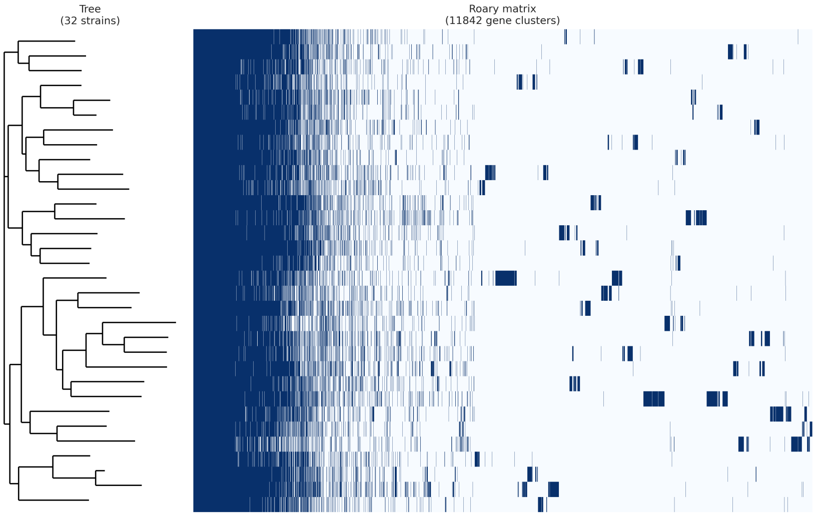


Figure 4. Figure showing a tree drawn considering the presence/absence of accessory genes on the left, obtained from the Roary matrix on the right.

This difference would keep increasing as more genomes are sequenced, which graphically would mean increasing the width of the Roary matrix in figure 4.

Taxonomic assignment

Taxonomy assignment was performed using PhyloPhlAn. All the MAGs belonging to our uSGB were taxonomically assigned to the *Lachnospiraceae* family, whereas the species remained unclassified, as it was not possible to find marker genes potentially able to assign the samples to different *geni*. From the analysis of one of the genomes we obtained the following output:

```
uSGB_4894:Genus:k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Lachnospiraceae|
                g__Lachnospiraceae_unclassified|s__Lachnospiraceae_unclassified_SGB4894|
                t__SGB4894:0.027455668421052626
```

All other genomes gave the same output except for the Mash distance, which oscillated around that value without ever going above 5% as expected.

Phylogenetic structure

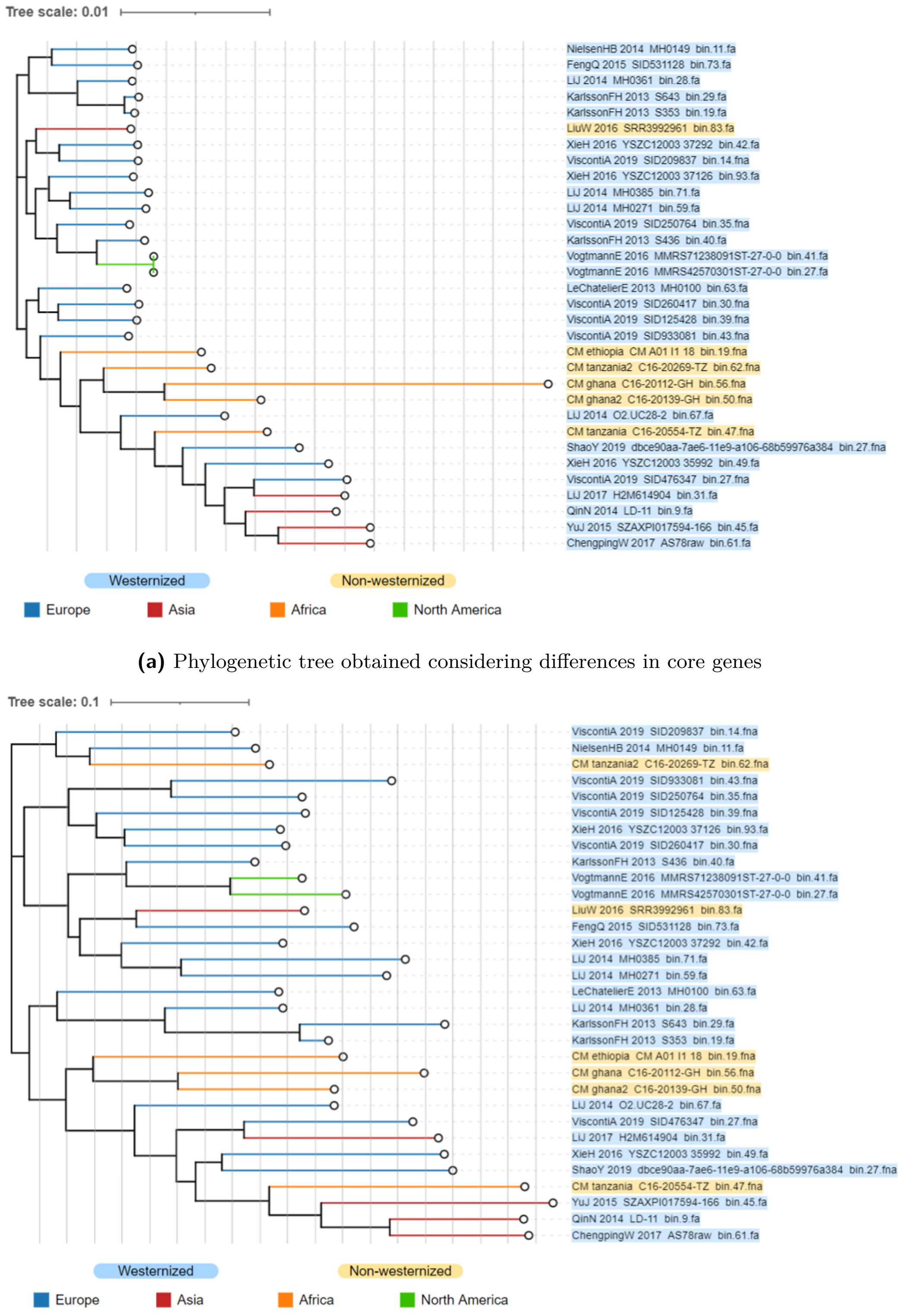
Phylogenetic maps were built using PhyloPhlAn 3.0, as described in the Method section. The results are shown in figure 5, where the first tree shows the distances between genomes calculated using differences in the core genes, while the second was built considering the presence/absence of accessory genes. With the aim of associating phylogeny with host data, we differentiated the branches on the basis of two criteria: *westernized/non-westernized* and *continent*. In both the graphs, we would suggest the presence of 3 different clusters derived from the root branch. This fact could signify a possible subdivision of our samples in different subspecies. Overall, no substantial differences were noticed among the two trees.

Taking into account the metadata of the genomes, from the tree we could infer a clustering role of the *continent* qualitative attribute. This type of correlation can be seen in both trees and it is more evident when considering exclusively the core genes. In fact, in the case of the core genome related graph, we noticed that all samples from Africa and Asia, except for one from Africa, belong to only one of the 3 clusters (third one from the top), whereas in the accessory genes tree all samples from Africa and Asia, except for 2 of them, belong to the third cluster. This indicates that this third putative strain (corresponding to the third cluster) might be more associated/present in the mentioned continents (Asia and Africa). Aside from the third branch, we did not notice any other significant geographical correlation. With regard to the *westernized/non-westernized* qualitative parameter instead, we did not observe any clear grouping of our MAGS.

Unfortunately, associations with other host data were unworkable due to scarcity of data (many missing values) or were not significant.

CONCLUSION

From the analysis of a uSGB, we firstly annotated and characterized the pangenome of the composing strains. We observed it to be open (figure 2-3), and computed the basic statistics of each MAG (table 1). Through taxonomic annotation, we found our genomes to belong to the *Lachnospiraceae* family, composed of obligately anaerobic, variably spore-forming bacteria[11]. We tried then to characterize phylogenetically the strains, and found that, in particular considering the differences in core genes(figure 5a), it was possible to observe a potential distinction in the MAGs through the *continent* parameter. As future perspectives, we would suggest further analysis of the Prokka generated data, for instance a more in depth functional analysis of the differentially expressed genes, and to get other samples from the Asian and the African continent. The latter recommendation would be needed to consolidate the clusters we saw and eventually to confirm them.


Tree scale: 0.01
NielsenHB 2014 MH0149 bin.11.fa
FengQ 2015 SID531128 bin.73.fa
LiJ 2014 MH0361 bin.28.fa
KarlssonFH 2013 S643 bin.29.fa
KarlssonFH 2013 S353 bin.19.fa
LiuW 2016 SRR3992961 bin.83.fa
XieH 2016 YSZC12003 37292 bin.42.fa
VisccontiA 2019 SID209837 bin.14.fna
XieH 2016 YSZC12003 37126 bin.93.fa
LiJ 2014 MH0385 bin.71.fa
LiJ 2014 MH0271 bin.59.fa
VisccontiA 2019 SID250764 bin.35.fna
KarlssonFH 2013 S436 bin.40.fa
VogtmannE 2016 MMRS71238091ST-27-0-0 bin.41.fa
VogtmannE 2016 MMRS42570301ST-27-0-0 bin.27.fa
LeChatelierE 2013 MH0100 bin.63.fa
VisccontiA 2019 SID260417 bin.30.fna
VisccontiA 2019 SID125428 bin.39.fna
VisccontiA 2019 SID933081 bin.43.fna
CM ethiopia CM A01 I1 18 bin.19.fna
CM tanzania2 C16-20269-TZ bin.62.fna
CM ghana C16-20112-GH bin.56.fna
CM ghana2 C16-20139-GH bin.50.fna
LiJ 2014 O2.UC28-2 bin.67.fa
CM tanzania C16-20554-TZ bin.47.fna
ShaoY 2019 dbce90aa-7ae6-11e9-a106-68b59976a384 bin.27.fna
XieH 2016 YSZC12003 35992 bin.49.fa
VisccontiA 2019 SID476347 bin.27.fna
LiJ 2017 H2M614904 bin.31.fa
QinN 2014 LD-11 bin.9.fa
YuJ 2015 SZAXPI017594-166 bin.45.fa
ChengpingW 2017 AS78raw bin.61.fa
Westernized
Non-westernized
Europe
Asia
Africa
North America

(a) Phylogenetic tree obtained considering differences in core genes

Tree scale: 0.1
VisccontiA 2019 SID209837 bin.14.fna
NielsenHB 2014 MH0149 bin.11.fa
CM tanzania2 C16-20269-TZ bin.62.fna
VisccontiA 2019 SID933081 bin.43.fna
VisccontiA 2019 SID250764 bin.35.fna
VisccontiA 2019 SID125428 bin.39.fna
XieH 2016 YSZC12003 37126 bin.93.fa
VisccontiA 2019 SID260417 bin.30.fna
KarlssonFH 2013 S436 bin.40.fa
VogtmannE 2016 MMRS71238091ST-27-0-0 bin.41.fa
VogtmannE 2016 MMRS42570301ST-27-0-0 bin.27.fa
LiuW 2016 SRR3992961 bin.83.fa
FengQ 2015 SID531128 bin.73.fa
XieH 2016 YSZC12003 37292 bin.42.fa
LiJ 2014 MH0385 bin.71.fa
LiJ 2014 MH0271 bin.59.fa
LeChatelierE 2013 MH0100 bin.63.fa
LiJ 2014 MH0361 bin.28.fa
KarlssonFH 2013 S643 bin.29.fa
KarlssonFH 2013 S353 bin.19.fa
CM ethiopia CM A01 I1 18 bin.19.fna
CM ghana C16-20112-GH bin.56.fna
CM ghana2 C16-20139-GH bin.50.fna
LiJ 2014 O2.UC28-2 bin.67.fa
VisccontiA 2019 SID476347 bin.27.fna
LiJ 2017 H2M614904 bin.31.fa
XieH 2016 YSZC12003 35992 bin.49.fa
ShaoY 2019 dbce90aa-7ae6-11e9-a106-68b59976a384 bin.27.fna
CM tanzania C16-20554-TZ bin.47.fna
YuJ 2015 SZAXPI017594-166 bin.45.fa
QinN 2014 LD-11 bin.9.fa
ChengpingW 2017 AS78raw bin.61.fa
Westernized
Non-westernized
Europe
Asia
Africa
North America

(b) Phylogenetic tree obtained considering accessory genes presence/absence

Figure 5. Phylogenetic trees obtained considering differences in the core genes (a) and accessory genes presence/absence (b)

References

1. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649–662.e20, DOI: 10.1016/j.cell.2019.01.001 (2019).

2. Segata Lab - Computational Metagenomics: Species-level genome bins (SGBs) from the human microbiome. http://segatalab.cibio.unitn.it/data/Pasolli_et_al.html.

3. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinforma. (Oxford, England)* **30**, 2068–2069, DOI: 10.1093/bioinformatics/btu153 (2014).

4. Andrew J. Page *et al.* Roary: Rapid large-scale prokaryote pan genome analysis (2015).

5. Andrew J. Page *et al.* R file from roary github repository. https://github.com/sanger-pathogens/Roary/blob/master/bin/create_pan_genome_plots.R.

6. Andrew J. Page *et al.* Python file from roary github repository. https://raw.githubusercontent.com/sanger-pathogens/Roary/master/contrib/roary_plots/roary_plots.py.

7. Löytynoja, A. Phylogeny-aware alignment with PRANK. *Methods Mol. Biol. (Clifton, N.J.)* **1079**, 155–170, DOI: 10.1007/978-1-62703-646-7_10 (2014).

8. Asnicar, F. *et al.* Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat. Commun.* **11**, 2500, DOI: 10.1038/s41467-020-16366-7 (2020).

9. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650, DOI: 10.1093/molbev/msp077 (2009).

10. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296, DOI: 10.1093/nar/gkab301 (2021).

11. Vacca, M. *et al.* The Controversial Role of Human Gut Lachnospiraceae. *Microorganisms* **8**, 573, DOI: 10.3390/microorganisms8040573 (2020).