Computational Microbial Genomics
Report
A.Y. 2021-2022

A Study to Identify Geographical Signatures in a Pangenome from Human Gut Microbiome

Submitted to: Prof. Dr. Nicola Segata
Submitted by: Surbhi Malhotra Surya Hembrom Annalisa Xamin

11 April 2022

University of Trento - Via Sommarive 9, 38123 Povo (TN), Italy

Contents

1 Introduction

Human gut microbiome is a complex environment, promoting several microbial species and strains survive and has been related to etiology of important human diseases across different populations and impacts overall human health[1, 2]. The gut of healthy and diseased individuals vary due to changes in gut microbiome and presence of various microbial species and strains and their differential functioning of genes [2]. Currently, Metagenome-assembled-genomes (MAGs) have led to rapid advancement in identification of various unprecedented gut microbial species and their strains with no or unfeasible standard lab-culturing techniques, or no high-quality reference genomes [3], even bypassing tedious lab-isolation and lab-culturing of hundreds of samples [4, 1] through affordable sequencing technologies [5]. The MAGs are constructed from contigs, which are formed by assembly of sequencing reads, through a binning process of single metagenome or several co-assembled metagenomes depending on nucleotide frequency, co-abundance, abundance (genes binned on co-abundance criteria are known as co-abundance gene groups (CAGs) [6]) and, or co-variation of abundance among several sample groups [1] presumed on k-mers[7]. The quality control of MAGs is crucial for recognition and removal of potential contaminants, identification of marker genes, and contiguity and completeness of metagenomes[1] before any comparative genomic analyses.
In this study we aimed to understand if human gut microbiome from different individuals exhibit any geographical signatures. For this, we examined high quality single taxon uSGB (unknown Species-level genome bin; SGB comprises either MAGs or MAGs and isolates aggregated from closely related strains of a species based on phylogenetics [5]) from individuals with no to varied health conditions from different countries. We also investigated the degree of within-taxon proximity of MAGs in this SGB to a known bacterial *Clostridium* species genome. For this, we determined the phylogenetic relationship of these SGBs to *Clostridium* spp. isolate. We even estimated, the proportions of core and accessory genes of all the SGBs to envisage any underlying pangenomic-level dynamics. Lastly, we investigated if these SGBs were conducive to specific disease-related pathogenicity in diseased or healthy individuals.

2 Materials and Methods

2.1 Sample data

We studied samples from project SGB6179 (uSGB) with 26 MAGs and 1 uncultured, whole genome shotgun sequenced *Clostridium* spp. isolate UMGS222. Out of which, 24 MAGs were human gut microbiome sampled from stools of 18 healthy, 3 diseased (1 with Type2 Diabetes: T2D, 1 with colorectal cancer: CRC, and 1 with HBV: Hepatitis B + HDV: Hepatitis D + cirrhosis) and 3 unknown health conditions. 2 MAGs were not described in metadata but are related to human stool sampled from gut microbiome studies by Nayfach et al., 2019 [1], and Nayfach et al., 2020[8]. Including healthy and unknown health conditions, we had 21 disease controls. All MAGs and isolate genome were checked for completeness of >90% and redundancy <5%. (see Appendix - Table 1, 2, 3).

2.2 Genome annotation

Genome annotation is labelling the CDS and intergenic regions inside the assembled genome. We annotated the MAGs and isolate fasta files with PROKKA, which incorporates several bioinformatics tools to acquire fast and reliable annotations of genomic bacterial sequences [9], with the following commands:

```
prokka --kingdom Bacteria --outdir prokka\_out --locustag L --prefix MAG\
    filename
```

wherein –kingdom represents the bacterial kingdom used for genome annotation, –locustag represents locustag, an identifier systematically attached to each gene. Following annotation, we extracted the number of CDS regions (in .txt files), hypothetical proteins and known protein (in .tsv files) per sample with custom bash scripts.

2.3 Pangenome analysis

Clustering of conspecific genomes with high confidence protein sequences with substantial amino acids identity engender pangenomes [10]. Pangenome analysis identifies the cumulative curve of genetic variability attributive of a given species with increase in individual genomes sequenced[11, 12]. We input the genome annotations (.gff files) from PROKKA into Roary[13] (using GNU parallel [14]) to retrieve microbial species' pangenome.
We did two runs of analyses with Roary:

1. Based on rapid Blastp alignment to check the presence or absence of the accessory genes.
2. Based on MAFFT[15] and PRANK[16] alignments of the core genes.

2.3.1 Based on Blastp alignments for presence or absence of accessory genes

We ran Roary with parameters -i for 95% identity cutoff for Blastp alignment (as 95% performed well, Figure 10) and -cd for 95% minimum threshold for all isolates to contain a gene to be classified as core gene, with default thread 1.

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

We processed the Roary alignment outputs for the presence/absence count of the accessory genes with Roary-inbuilt R-enabled python script roary_plots.py (`https://raw.githubusercontent.com/sanger-pathogens/Roary/master/contrib/roary_plots/roary_plots.py`) with the following commands:

```
python3 roary_plots.py accessory_binary_genes.fa.newick gene_presence_absence.csv
```

We obtained:

1. Pangenome frequency plot
2. Presence and absence matrix plot against the tree
3. Pangenome pie-chart (core, soft core, shell and cloud genes)

We used Roary-inbuilt R script create_pan_genome_plots.R `https://github.com/sanger-pathogens/Roary/blob/master/bin/create_pan_genome_plots.R` to understand the dynamics of pangenome. We retrieved the total number of genes under four different categories i.e., core, soft, shell, and cloud genes forming the pangenome.
We obtained plots with:

1. The number of Blastp hits with different percentage identity.
2. The number of conserved and total genes with increase in the number of genomes.
3. The number of unique and new genes with increase in the number of genomes.

2.3.2 Based on alignments of the core genes

We ran Roary for multiFASTA alignment of core genes[13]. with additional parameters -e for slow and accurate alignment with inbuilt PRANK and -n for fast alignment with inbuilt MAFFT using default thread 1.

```
roary *.gff -f roary_out_align -e -n -i 95 -cd 95
```

We did further alignment analyses using the same create_pan_genome_plots.R and roary_plots.py scripts.

We compared the results from first run of Roary to its second run the second analysis. These results are eminent for downstream analyses such as phylogenetic tree reconstruction and SNPs identification[17].

2.4 Taxonomic characterisation

We taxonomically characterised the MAGs with PhyloPhlAn[5].

```
phylophlan_metagenomic -i phylophlan_input -o phylophlan_output --nproc 4 -n 1
    --database_update -d CMG2122 --verbose -e .fa
```

wherein we used parameters –nproc for 4 number of CPUs; -n for number of best hit within each MAG matching the database to retain, -d for database CMG2122 with –database_update for database updation, -e for fasta format (.fna or .fa).

2.5 Phylogenetic structure

We visualized the phylogenetic trees from MAGs and isolate with Interactive Tree Of Life(iTOL) v6, an online tool for the display, annotation and management of phylogenetic and other trees[18]. We reconstructed phylogenetic trees from two different Roary analyses:

1. **Phylogenetic tree based on presence/absence of accessory genes (Roary)**: Uploaded accessory_binary_genes.fa.newick to the iTOL.

2. **Phylogenetic tree based on core genes alignment (Roary with additional -e and -n parameters)**: Using FastTree[19] with -nt for nucleotide, we generated a phylogenetic tree core_gene.tre from core_gene_alignment.aln and visualized in iTOL.

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

We used R[20] for generation of high quality statistical plots.

3 Results and Discussion

We processed high quality MAGs with contigs ranging from 71 to 449. The least number of contigs were in MAGs: ShaoY_2019___cc7b0cfa-7ae6-11e9-a106-68b59976a384___bin.21 (71 contigs), GCA_900540255 (79 contigs), QinJ_2012___T2D-014___bin.33 (72 contigs) to as high as CM_Neuroblastoma___NB_CTR79___bin.26 (449 contigs), ViscontiA_2019___SID129237___bin.45 (446 contigs). We found no substantial relationship between completeness of MAGs to number of contigs and redundancy of MAGs to number of contigs (see Figures 8,9). The CDS counts per MAGs were approx. 2500 to 3000s. The hypothetical proteins ranged between approx. 1000 to 1330. It was relatively low for CM_guinea2___GUI_90404___bin.43 (901), CM_guinea___GUI_0080302___bin.8 (844). The known proteins were in range of approx. 1400 to 1500s (see Appendix - Table 4).

The variation in contigs number per MAGs, hypothetical proteins, known proteins could indicate the richness of microbiome in some individuals than the rest and the MAGs are of high quality.

3.1 Roary: based on Blastp alignment for presence/absence of genes

We observed a nearly linear increase in total genes whereas exponential decrease inconsistently to a constant plateau in conserved genes with increasing number of genomes (MAGs and isolate) (Figure 1). The number of unique genes grow exponentially with increasing number of genomes (Figure 2). With increasing number of MAGs, the new genes' number decreased exponentially in an inconsistent manner with sudden peaks and drops. Studies reveal that the total size of a pangenome stabilizes eventually (i.e., plateau formation in an initially exponential curve) is typical of closed pangenomes [21][22]. Since we did not observe such plateau for new and unique genes, thus this uSGB forms an open pangenome. However, to re-establish our findings, more genomes need to be analysed to estimate the total genetic complement of this species and understand the evolutionary dynamics.

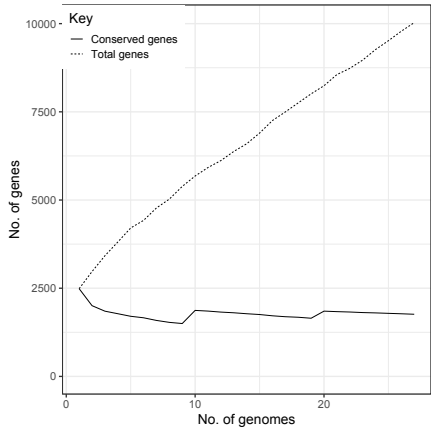

Figure 1: Conserved genes and total genes across pangenome

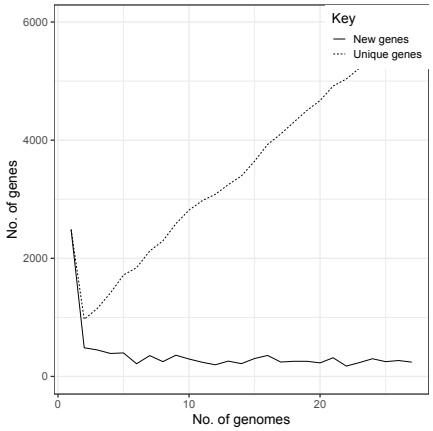

Figure 2: New genes and unique genes across pangenome

We observed a small proportion of core genes present in the 27 genomes/strains whereas a large proportion of genes is specific to single or few genomes. Genes specific to few genomes could be newly acquired genes or unique genes of the given genome (Figure 3). We obtained a total of 10028 genes, wherein 1036 were core genes, 513 soft-core genes, 1180 shell genes and 7299 were cloud genes (Figure 4). Through the heatmap we obtained the presence and absence of 10028 genes. We found that only approx. one-tenth of total genes are present in all the strains, i.e. are core genes. Majority of genes are not present in all the strains. This could indicate that this pangenome has lesser core genes, and many accessory genes are strain-specific (Figure 5).

3.2 Roary: using alignment with MAFFT and PRANK

After the core genes alignment, the number of Blastp hits under different percentage identity (Figure 11), increase in total and unique genes, decrease in conserved genes and new genes per total number of genomes (Figures 12,13), frequency of genes across pangenome (Figure 14) showed trend similar to previous run of Roary (Figures 10,1,2,3). However, in the second run of Roary, a total 10020 genes with 1035 core genes, 513 soft-core genes, 1181 shell genes and 7291 cloud genes (Figure 15) were found and the presence/absence of genes (Figure 16) varied as well (Figures 4, 5).

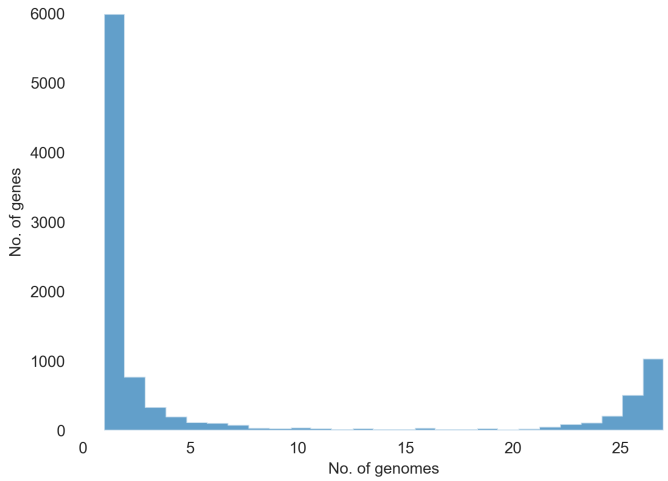

Figure 3: Frequency of genes across pangenome

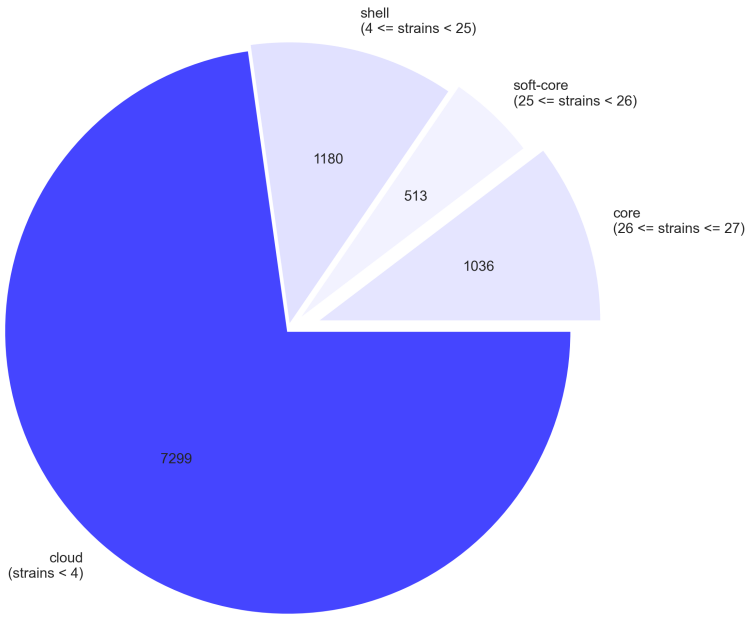

Figure 4: Pangenome genes composition: cloud (7299), shell (1180), soft-core (513) and core (1036) genes

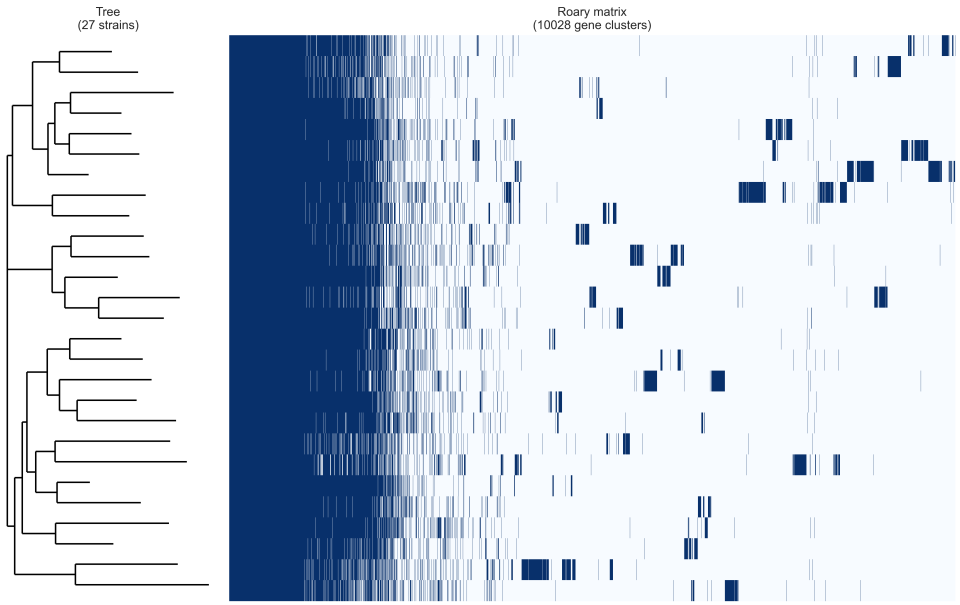

Figure 5: Heatmap of pangenome. Dark blue represents gene presence; light blue represents gene absence.The x-axis represents 10028 genes clusters and y-axis represents 27 strains in dendrogram. The gene clusters at left in dark blue depicts core genes.

This is due to differences in the working of alignment software used by Roary, thus indicating Blastp implementation is less sensitive than the PRANK and MAFFT in finding core and accessory genes.

3.3 Phylogenetic structure

From the phylogeny of aligned core genes in the pangenome, we observed three superclades with strains from individuals from different geographical locations clustered together (Figures 6,1). The topmost superclade had ViscontiA_2019SID129237 (Great Britain, GBR), LiJ_2014V1.UC22-1 (Spain) and XieH_2016YSZC12003_35705 (GBR) clustered. QinN_2014LD-41 (China; HBV+HDV+Cirrhosis), QinJ_2012T2D-014 (China; T2D), ShaoY_2019a504a8ac-7ae6-11e9-a106-68b59976a384 (GBR), ShaoY_2019-_ SID815390bc-7ae6-11e9-a106-68b59976a384 (GBR), YuJ_2015SZAXPI015233-19 (CHN), YuJ_2015SZAXPI015252-43 (CHN), YuJ_2015SZAXPI003424-12 (CHN), PasolliE_2018 _Madagascar A14_01_1FE_CM_MDG_14011 (Madagascar) and GCA900540255 (isolate), NayfachS_2020_GEM_3300029556 (unknown), CM_Guinea GUI_80104 (Guinea) clustered together. The other strains clustered in the bottom-most superclade were 1 from Italy (CM_Neuroblastoma_NB_CTR79), 3 from GBR (ShaoY_2019 b3923042-7ae6-11e9-a106-68b59976a384, ShaoY_2019 afafe9a6-7ae6-11e9-a106-68b59976a384, ShaoY_2019cc7b0cfa-7ae6-11e9-a106-68b59976a384) and rest 8 from Guinea (CM guinea2 strains). This indicates that the core genes from all strains express uniformly in guts of different individuals from different geographical locations and might indicate that they undergo balanced selection. The gut microbial strains of individuals with comorbidities or single disease clustered freely with those of healthy individuals and *Clostridium* isolate, hence, suggest little indication of these strains in causation of these diseases. However, we need to re-affirm this fact by examining more diseased individuals. The gut microbiome of T2D and CRC patients has elevated diversity of species from phylum Firmicutes, class

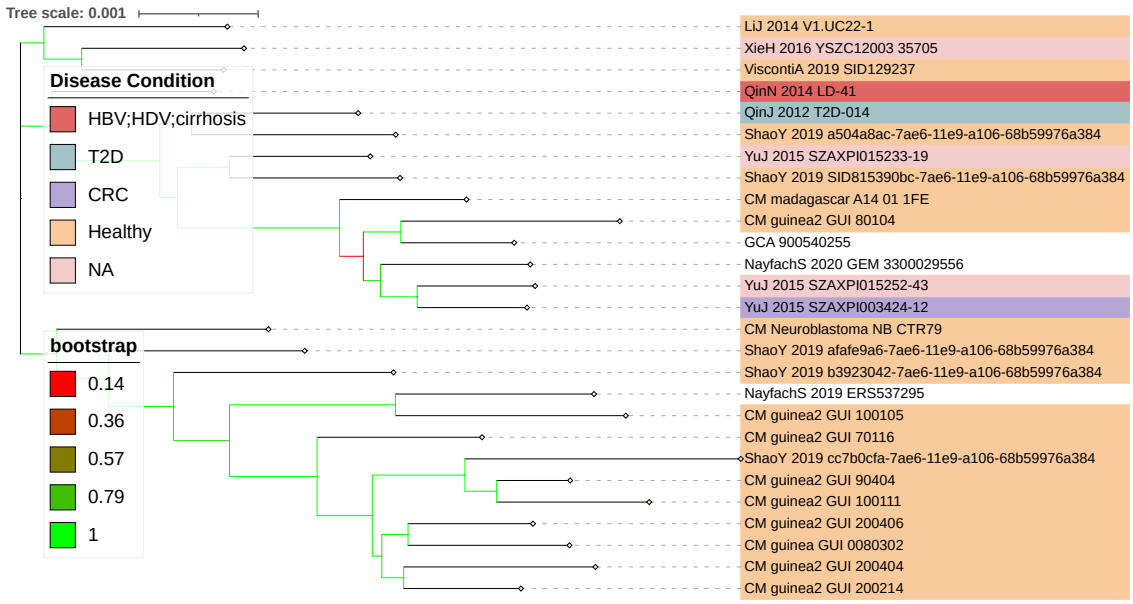

Figure 6: Phylogenetic tree through core genes alignment. The gut microbiome MAGs from healthy individuals are indicated in yellow, diseased patient with HBV (Hepatitis-V) + HDV (Hepatitis-D) + Liver cirrhosis in red, T2D (Type 2 Diabetes) in green and CRC (Colorectal Cancer) in purple. The *Clostridium* isolate and NayfachS samples in white and individuals with unknown health condition in pink.

Clostridia whereas liver cirrhosis patients have similar overall microbiome representation as in healthy ones [23, 24, 1, 2]. Further research is needed to understand the *Clostridium*-disease pathogenic associations.

From the phylogeny of accessory genes in the pangenome, based on genes' presence/absence, we detected some patterns of microbiome diversity among individuals from different geographical locations (Figures 7,1). Four samples from- China (2; 1 unknown health condition and 1 with CRC), Madagascar (1; healthy), Guinea (1; healthy) clustered with the isolate GCA_900540255 in the first superclade. We noticed that rest 8 samples from Guinea (healthy and non-westernised (NW) except CM_guinea_ GUI0080302- westernized) clustered together along with NayfachS_2019 sample unlike their clustering of their core genes (Figures 6, 1). The remaining samples from- GBR (5; healthy), Italy (1; healthy), Spain (1; healthy), China (4; 2 diseased and 2 unknown health condition), all from westernized populations, clustered together. This confers that the geographical location and lifestyle conditions govern the fixation/loss of accessory genes of the *Clostridium* spp. in the human gut of individuals from a given geo-location. We recognised that the core and accessory genes of the *Clostridium* isolate is closer to samples from Guinea (NW), Madagascar (NW) and China (westernized) than to GBR, Spain or Italy.

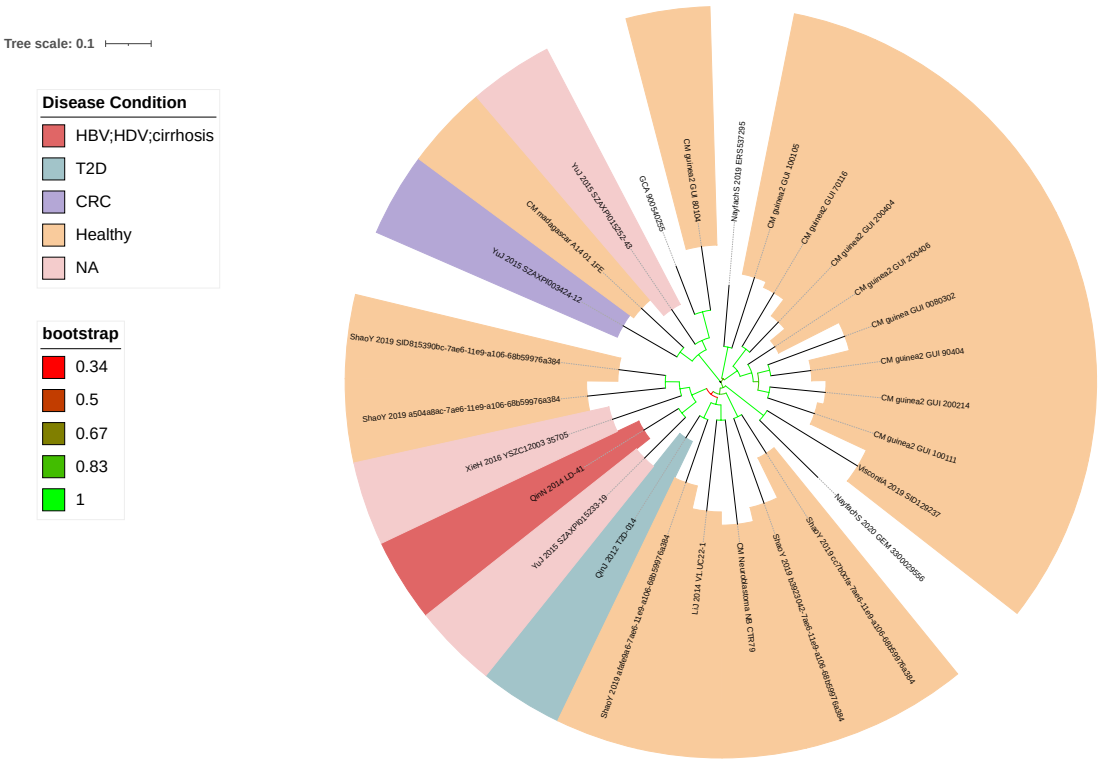

Figure 7: Phylogenetic tree of accessory genes indicating their presence/absence. The gut microbiome MAGs from healthy individuals are indicated in yellow, diseased patient with HBV (Hepatitis-V) + HDV (Hepatitis-D) + Liver cirrhosis in red, T2D (Type 2 Diabetes) in green and CRC (Colorectal Cancer) in purple. The *Clostridium* isolate and NayfachS samples in white and individuals with unknown health condition in pink.

3.4 Taxonomic characterisation

We noticed that all MAGs and the isolate had <0.05 as taxonomical distances, which means strong similarities among the different MAGs. We observed that each human gut microbiome represents a different strain of the same species *Clostridium SGB6179*. The phylogenetic richness is typical of Firmicutes, class Clostridia [25] hence, the phylogenetic tree in our study even corroborates this fact that, Clostridia has open pangenome, possibly due to lateral gene transfer among strains [23].

The MAGs contribute to functional characterisation and taxonomical classification of unfamiliar less copious human gut microbial species and their potential role in pathogenicity and physiology inside host human gut[3, 10]. The construction of MAGs from metagenomes databases could diagnose the genetic distinctness and novelty of human gut microbiome among populations worldwide [1, 5] which we witnessed in this study. Such studies can accelerate disease diagnostics and possible cure for human diseases and characterise microbes within the microbiomes which incur diseases non-specifically or specifically as discussed in [2]. We found that the accessory genes could be population-specific corroborating to recent findings[10]. Crucial metabolic pathways and many housekeeping genes-related pathways are driven by core genes (i.e. genes found in more than 90% of conspecific genomes). Contrarily, accessory genes (i.e. genes found in less than 10% of conspecific genomes) regulate recombination, replication, comprise mobile genetic elements and control defense and resistance machinery[10]. We found that this uSGB belongs to *Clostridium* spp. and according to some studies[26], that a single strain colonises adult human gut, it would be meaningful to identify such predominant strains and their level of the pathogenicity among different populations for this uSGB.

4 Conclusions

We studied that different MAGs inside uSGB belong to *Clostridium* spp. We even found several new and unique genes in the pangenome constructed from these MAGs. We found that the pangenome is open in nature, i.e. acquiring new genes gradually. The core genes are roughly conserved among all MAGs despite the samples were obtained from different geographical locations. But, we noticed that the accessory genes are fixing in the various populations and becoming population-specific, inferring that the geography and lifestyle plays a role in such a phenomenon.

5 References

References

[1] S. Nayfach, Z. J. Shi, R. Seshadri, K. S. Pollard, and N. C. Kyrpides, "New insights from uncultivated genomes of the global human gut microbiome," *Nature*, vol. 568, no. 7753, pp. 505–510, Mar. 2019. [Online]. Available: https://doi.org/10.1038/s41586-019-1058-x

[2] C. R. Armour, S. Nayfach, K. S. Pollard, and T. J. Sharpton, "A metagenomic meta-analysis reveals functional signatures of health and disease in the human gut microbiome," *mSystems*, vol. 4, no. 4, Aug. 2019. [Online]. Available: https://doi.org/10.1128/msystems.00332-18

[3] H. Jin, L. You, F. Zhao, S. Li, T. Ma, L.-Y. Kwok, H. Xu, and Z. Sun, "Hybrid, ultra-deep metagenomic sequencing enables genomic and functional characterization of low-abundance species in the human gut microbiome," *Gut Microbes*, vol. 14, no. 1, Jan. 2022. [Online]. Available: https://doi.org/10.1080/19490976.2021.2021790

[4] M. Watson, "New insights from 33,813 publicly available metagenome-assembled-genomes (MAGs) assembled from the rumen microbiome," *bioRxiv*, 2021. [Online]. Available: https://www.biorxiv.org/content/early/2021/04/02/2021.04.02.438222

[5] F. Asnicar, A. M. Thomas, F. Beghini, C. Mengoni, S. Manara, P. Manghi, Q. Zhu, M. Bolzan, F. Cumbo, U. May, J. G. Sanders, M. Zolfo, E. Kopylova, E. Pasolli, R. Knight, S. Mirarab, C. Huttenhower, and N. Segata, "Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0," *Nature Communications*, vol. 11, no. 1, May 2020. [Online]. Available: https://doi.org/10.1038/s41467-020-16366-7

[6] H. B. Nielsen, , M. Almeida, A. S. Juncker, S. Rasmussen, J. Li, S. Sunagawa, D. R. Plichta, L. Gautier, A. G. Pedersen, E. L. Chatelier, E. Pelletier, I. Bonde, T. Nielsen, C. Manichanh, M. Arumugam, J.-M. Batto, M. B. Q. dos Santos, N. Blom, N. Borruel, K. S. Burgdorf, F. Boumezbeur, F. Casellas, J. Doré, P. Dworzynski, F. Guarner, T. Hansen, F. Hildebrand, R. S. Kaas, S. Kennedy, K. Kristiansen, J. R. Kultima, P. Léonard, F. Levenez, O. Lund, B. Moumen, D. L. Paslier, N. Pons, O. Pedersen, E. Prifti, J. Qin, J. Raes, S. Sørensen, J. Tap, S. Tims, D. W. Ussery, T. Yamada, P. Renault, T. Sicheritz-Ponten, P. Bork, J. Wang, S. Brunak, and S. D. Ehrlich, "Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes," *Nature Biotechnology*, vol. 32, no. 8, pp. 822–828, Jul. 2014. [Online]. Available: https://doi.org/10.1038/nbt.2939

[7] J. C. Setubal, "Metagenome-assembled genomes: concepts, analogies, and challenges," *Biophysical Reviews*, vol. 13, no. 6, pp. 905–909, Nov. 2021. [Online]. Available: https://doi.org/10.1007/s12551-021-00865-y

[8] Z. J. Shi, B. Dimitrov, C. Zhao, S. Nayfach, and K. S. Pollard, "Ultra-rapid metagenotyping of the human gut microbiome," Jun. 2020. [Online]. Available: https://doi.org/10.1101/2020.06.12.149336

[9] T. Seemann, "Prokka: rapid prokaryotic genome annotation," *Bioinformatics*, vol. 30, no. 14, pp. 2068–2069, Mar. 2014. [Online]. Available: https://doi.org/10.1093/bioinformatics/btu153

[10] A. Almeida, S. Nayfach, M. Boland, F. Strozzi, M. Beracochea, Z. J. Shi, K. S. Pollard, D. H. Parks, P. Hugenholtz, N. Segata, N. C. Kyrpides, and R. D. Finn, "A

unified sequence catalogue of over 280,000 genomes obtained from the human gut microbiome," Sep. 2019. [Online]. Available: https://doi.org/10.1101/762682

[11] A. Muzzi and C. Donati, "Population genetics and evolution of the pan-genome of *Streptococcus pneumoniae*," *International Journal of Medical Microbiology*, vol. 301, no. 8, pp. 619–622, 2011, 9th International Meeting on Microbial Epidemiological Markers. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1438422111000944

[12] N. Nguyen, G. Hickey, D. R. Zerbino, B. Raney, D. Earl, J. Armstrong, W. J. Kent, D. Haussler, and B. Paten, "Building a pan-genome reference for a population," *Journal of Computational Biology*, vol. 22, no. 5, pp. 387–401, 2015, pMID: 25565268. [Online]. Available: https://doi.org/10.1089/cmb.2014.0146

[13] A. J. Page, C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. Holden, M. Fookes, D. Falush, J. A. Keane, and J. Parkhill, "Roary: rapid large-scale prokaryote pan genome analysis," *Bioinformatics*, vol. 31, no. 22, pp. 3691–3693, Jul. 2015. [Online]. Available: https://doi.org/10.1093/bioinformatics/btv421

[14] O. Tange, "Gnu parallel 20220222 ('donetsk luhansk')," Feb. 2021, GNU Parallel is a general parallelizer to run multiple serial command line programs in parallel without changing them. [Online]. Available: https://doi.org/10.5281/zenodo.6213471

[15] K. Katoh, "MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform," *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, Jul. 2002. [Online]. Available: https://doi.org/10.1093/nar/gkf436

[16] A. Löytynoja, "Phylogeny-aware alignment with PRANK," in *Methods in Molecular Biology*. Humana Press, Aug. 2013, pp. 155–170. [Online]. Available: https://doi.org/10.1007/978-1-62703-646-7_10

[17] F. Sitto and F. U. Battistuzzi, "Estimating Pangenomes with Roary," *Molecular Biology and Evolution*, vol. 37, no. 3, pp. 933–939, 12 2019. [Online]. Available: https://doi.org/10.1093/molbev/msz284

[18] I. Letunic and P. Bork, "Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation," *Nucleic Acids Research*, vol. 49, no. W1, pp. W293–W296, Apr. 2021. [Online]. Available: https://doi.org/10.1093/nar/gkab301

[19] M. N. Price, P. S. Dehal, and A. P. Arkin, "FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix," *Molecular Biology and Evolution*, vol. 26, no. 7, pp. 1641–1650, Apr. 2009. [Online]. Available: https://doi.org/10.1093/molbev/msp077

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[21] H. Tettelin, V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. DeBoy, T. M. Davidsen, M. Mora, M. Scarselli, I. M. y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. B. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser, "Genome analysis of multiple pathogenic isolates of *Streptococcus*

agalactiae: Implications for the microbial pangenome;," *Proceedings of the National Academy of Sciences*, vol. 102, no. 39, pp. 13 950–13 955, 2005. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.0506758102

[22] H. Tettelin, D. Riley, C. Cattuto, and D. Medini, "Comparative genomics: the bacterial pan-genome," *Current Opinion in Microbiology*, vol. 11, no. 5, pp. 472–477, 2008, antimicrobials/Genomics. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1369527408001239

[23] J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, Y. Peng, D. Zhang, Z. Jie, W. Wu, Y. Qin, W. Xue, J. Li, L. Han, D. Lu, P. Wu, Y. Dai, X. Sun, Z. Li, A. Tang, S. Zhong, X. Li, W. Chen, R. Xu, M. Wang, Q. Feng, M. Gong, J. Yu, Y. Zhang, M. Zhang, T. Hansen, G. Sanchez, J. Raes, G. Falony, S. Okuda, M. Almeida, E. LeChatelier, P. Renault, N. Pons, J.-M. Batto, Z. Zhang, H. Chen, R. Yang, W. Zheng, S. Li, H. Yang, J. Wang, S. D. Ehrlich, R. Nielsen, O. Pedersen, K. Kristiansen, and J. Wang, "A metagenome-wide association study of gut microbiota in type 2 diabetes," *Nature*, vol. 490, no. 7418, pp. 55–60, Sep. 2012. [Online]. Available: https://doi.org/10.1038/nature11450

[24] N. Qin, F. Yang, A. Li, E. Prifti, Y. Chen, L. Shao, J. Guo, E. L. Chatelier, J. Yao, L. Wu, J. Zhou, S. Ni, L. Liu, N. Pons, J. M. Batto, S. P. Kennedy, P. Leonard, C. Yuan, W. Ding, Y. Chen, X. Hu, B. Zheng, G. Qian, W. Xu, S. D. Ehrlich, S. Zheng, and L. Li, "Alterations of the human gut microbiome in liver cirrhosis," *Nature*, vol. 513, no. 7516, pp. 59–64, Jul. 2014. [Online]. Available: https://doi.org/10.1038/nature13568

[25] S. C. Forster, N. Kumar, B. O. Anonye, A. Almeida, E. Viciani, M. D. Stares, M. Dunn, T. T. Mkandawire, A. Zhu, Y. Shao, L. J. Pike, T. Louie, H. P. Browne, A. L. Mitchell, B. A. Neville, R. D. Finn, and T. D. Lawley, "A human gut bacterial genome and culture collection for improved metagenomic analyses," *Nature Biotechnology*, vol. 37, no. 2, pp. 186–192, Feb. 2019. [Online]. Available: https://doi.org/10.1038/s41587-018-0009-7

[26] P. Ferretti, E. Pasolli, A. Tett, F. Asnicar, V. Gorfer, S. Fedi, F. Armanini, D. T. Truong, S. Manara, M. Zolfo, F. Beghini, R. Bertorelli, V. D. Sanctis, I. Bariletti, R. Canto, R. Clementi, M. Cologna, T. Crifò, G. Cusumano, S. Gottardi, C. Innamorati, C. Masè, D. Postai, D. Savoi, S. Duranti, G. A. Lugli, L. Mancabelli, F. Turroni, C. Ferrario, C. Milani, M. Mangifesta, R. Anzalone, A. Viappiani, M. Yassour, H. Vlamakis, R. Xavier, C. M. Collado, O. Koren, S. Tateo, M. Soffiati, A. Pedrotti, M. Ventura, C. Huttenhower, P. Bork, and N. Segata, "Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome," *Cell Host & Microbe*, vol. 24, no. 1, pp. 133–145.e5, Jul. 2018. [Online]. Available: https://doi.org/10.1016/j.chom.2018.06.005

6 Appendix

Dataset name	Sample ID	Number of reads	Number of bases	Minimum read length	Median read length
LiJ_2014	V1.UC22-1	82978692	6823277690	30	86
QinJ_2012	T2D-014	52225992	4700339280	90	90
QinN_2014	LD-41	52250886	5225088600	100	100
XieH_2016	YSZC12003_35705	66991690	6564082392	30	100
YuJ_2015	SZAXPI003424-12	58546892	5593003062	30	100
YuJ_2015	SZAXPI015233-19	72077438	7024760687	30	100
YuJ_2015	SZAXPI015252-43	44904362	4378527165	30	100
CM_Guinea	GUI_0080302	37003882	3718239745	75	101
CM_Guinea	GUI_100105	49958829	7499685933	75	151
CM_Guinea	GUI_100111	46045645	6918322696	75	151
CM_Guinea	GUI_200214	48425161	7270574229	75	151
CM_Guinea	GUI_200404	61377875	9213316332	75	151
CM_Guinea	GUI_200406	49968611	7507455459	75	151
CM_Guinea	GUI_70116	46373584	6970616703	75	151
CM_Guinea	GUI_80104	62630186	9407192009	75	151
CM_Guinea	GUI_90404	23708901	3556954184	75	151
CM_NEUROBLASTOMA	NB_CTR79	54826722	8235060533	75	151
PasolliE_2018_Madagascar	A14_01_1FE	51877496	5117619802	75	101
ShaoY_2019	a504a8ac-7ae6-11e9-a106-68b59976a384	13388070	1646174990	NA	125
ShaoY_2019	cc7b0cfa-7ae6-11e9-a106-68b59976a384	16934168	2009984998	NA	125
ShaoY_2019	b3923042-7ae6-11e9-a106-68b59976a384	18601446	2235383168	NA	125
ShaoY_2019	afafe9a6-7ae6-11e9-a106-68b59976a384	16019068	1918265711	NA	125
ShaoY_2019	SID815390bc-7ae6-11e9-a106-68b59976a384	19783784	2378314506	NA	125
ViscontiA_2019	SID129237	50085124	5839317670	75	124

Table 1: SGB6179 metadata: Moreover, in our dataset, we had also a genome coming from known isolated organism, *Clostridium sp.*. This genome is present in our dataset as GCA_900540255.fna.

Dataset name	Subject ID	Study conditions	Disease	Age	Gender	Country	Non Westernized
LiJ_2014	V1.UC22-1	control	healthy	NA	NA	ESP	no
QinJ_2012	T2D-014	T2D	T2D	63	female	CHN	no
QinN_2014	LD-41	cirrhosis	HBV;HDV;cirrhosis	47	female	CHN	no
XieH_2016	YSZC12003_35705	control	NA	68	female	GBR	no
YuJ_2015	SZAXPI003424-12	CRC	CRC	NA	NA	CHN	no
YuJ_2015	SZAXPI015233-19	control	NA	NA	NA	CHN	no
YuJ_2015	SZAXPI015252-43	control	NA	NA	NA	CHN	no
CM_Guinea	GUI_0080302	control	healthy	24	female	GUI	no
CM_Guinea	GUI_100105	control	healthy	6	female	GUI	yes
CM_Guinea	GUI_100111	control	healthy	6	female	GUI	yes
CM_Guinea	GUI_200214	control	healthy	NA	NA	GUI	yes
CM_Guinea	GUI_200404	control	healthy	20	female	GUI	yes
CM_Guinea	GUI_200406	control	healthy	16	female	GUI	yes
CM_Guinea	GUI_70116	control	healthy	45	female	GUI	yes
CM_Guinea	GUI_80104	control	healthy	8	male	GUI	yes
CM_Guinea	GUI_90404	control	healthy	4	female	GUI	yes
CM_NEUROBLASTOMA	NB_CTR79	control	healthy	5.8	male	ITA	N
PasolliE_2018_Madagascar	CM_MDG_14011	control	healthy	36	male	MDG	yes
ShaoY_2019	B01339	control	healthy	0.010958904	male	GBR	no
ShaoY_2019	B02739	control	healthy	0.575342466	male	GBR	no
ShaoY_2019	B01799	control	healthy	0.769863014	female	GBR	no
ShaoY_2019	B01712	control	healthy	0.8	male	GBR	no
ShaoY_2019	A01685	control	healthy	0	male	GBR	no
ViscontiA_2019	TUK89005992	control	healthy	60	female	GBR	no

Table 2: SGB6179 metadata

Bin	Sample	Completeness	Redundancy	SGB
CM_guinea2___GUI_100105___bin.75	GUI_100105	94.35	2.67	SGB6179
CM_guinea2___GUI_100111___bin.34	GUI_100111	95.97	0.81	SGB6179
CM_guinea2___GUI_200214___bin.86	GUI_200214	95.97	1.45	SGB6179
CM_guinea2___GUI_200404___bin.54	GUI_200404	92.99	3.02	SGB6179
CM_guinea2___GUI_200406___bin.2	GUI_200406	95.97	1.05	SGB6179
CM_guinea2___GUI_70116___bin.90	GUI_70116	92.79	1.33	SGB6179
CM_guinea2___GUI_80104___bin.57	GUI_80104	95.65	4.87	SGB6179
CM_guinea2___GUI_90404___bin.43	GUI_90404	95.71	3.92	SGB6179
CM_guinea___GUI_0080302___bin.8	GUI_0080302	92.1	1.37	SGB6179
CM_madagascar___A14_01_1FE___bin.13	A14_01_1FE	95.96774194	0.161290323	SGB6179
CM_Neuroblastoma___NB_CTR79___bin.26	NB_CTR79	93.86	3.5	SGB6179
GCA_900540255	SAMEA4890906	96.77	0	SGB6179
LiJ_2014___V1.UC22-1___bin.24	V1.UC22-1	91.46505376	2.459677419	SGB6179
NayfachS_2019___ERS537295___bin.57	ERS537295	95.16	0	SGB6179
NayfachS_2020___GEM_3300029556___bin.6	GEM_3300029556	92.34	4.34	SGB6179
QinJ_2012___T2D-014___bin.33	T2D-014	96.77419355	0	SGB6179
QinN_2014___LD-41___bin.25	LD-41	94.40092166	0.089605735	SGB6179
ShaoY_2019___a504a8ac-7ae6-11e9-a106-68b59976a384___bin.8	a504a8ac-7ae6-11e9-a106-68b59976a384	95.57	0.43	SGB6179
ShaoY_2019___afafe9a6-7ae6-11e9-a106-68b59976a384___bin.6	afafe9a6-7ae6-11e9-a106-68b59976a384	94.35	0.81	SGB6179
ShaoY_2019___b3923042-7ae6-11e9-a106-68b59976a384___bin.23	b3923042-7ae6-11e9-a106-68b59976a384	95.16	1.47	SGB6179
ShaoY_2019___cc7b0cfa-7ae6-11e9-a106-68b59976a384___bin.21	cc7b0cfa-7ae6-11e9-a106-68b59976a384	95.16	0.16	SGB6179
ShaoY_2019___SID815390bc-7ae6-11e9-a106-68b59976a384___bin.19	SID815390bc-7ae6-11e9-a106-68b59976a384	95.97	0.16	SGB6179
ViscontiA_2019___SID129237___bin.45	SID129237	91.29	0	SGB6179
XieH_2016___YSZC12003_35705___bin.27	YSZC12003_35705	95.11776754	1.500896057	SGB6179
YuJ_2015___SZAXPI003424-12___bin.14	SZAXPI003424-12	95.88709677	2.777777778	SGB6179
YuJ_2015___SZAXPI015233-19___bin.67	SZAXPI015233-19	96.77419355	0.403225806	SGB6179
YuJ_2015___SZAXPI015252-43___bin.58	SZAXPI015252-43	96.77419355	1.792114695	SGB6179

Table 3: SGB6179 bin data

Samples	Number of contigs	CDS counts	Hypothetical proteins	Known proteins
CM_Neuroblastoma___NB_CTR79___bin.26	449	2619	1231	1458
CM_guinea2___GUI_100105___bin.75	233	3192	1718	1549
CM_guinea2___GUI_100111___bin.34	109	2688	1238	1540
CM_guinea2___GUI_200214___bin.86	282	2765	1313	1525
CM_guinea2___GUI_200404___bin.54	261	2367	1027	1388
CM_guinea2___GUI_200406___bin.2	223	2900	1401	1589
CM_guinea2___GUI_70116___bin.90	339	2307	1023	1360
CM_guinea2___GUI_80104___bin.57	306	2583	1142	1516
CM_guinea2___GUI_90404___bin.43	145	2244	901	1412
CM_guinea___GUI_0080302___bin.8	261	2170	844	1380
CM_madagascar___A14_01_1FE___bin.13	261	2536	1124	1447
GCA_900540255	79	2568	1138	1465
LiJ_2014___V1.UC22-1___bin.24	337	2394	1036	1383
NayfachS_2019___ERS537295___bin.57	150	2522	1110	1467
NayfachS_2020___GEM_3300029556___bin.6	282	2862	1334	1594
QinJ_2012___T2D-014___bin.33	72	2467	1032	1513
QinN_2014___LD-41___bin.25	90	2471	1062	1473
ShaoY_2019___SID815390bc-7ae6-11e9-a106-68b59976a384___bin.19	86	2570	1152	1487
ShaoY_2019___a504a8ac-7ae6-11e9-a106-68b59976a384___bin.8	173	2531	1113	1492
ShaoY_2019___afafe9a6-7ae6-11e9-a106-68b59976a384___bin.6	100	2565	1139	1499
ShaoY_2019___b3923042-7ae6-11e9-a106-68b59976a384___bin.23	154	2632	1176	1530
ShaoY_2019___cc7b0cfa-7ae6-11e9-a106-68b59976a384___bin.21	71	2495	1101	1475
ViscontiA_2019___SID129237___bin.45	446	2438	1058	1431
XieH_2016___YSZC12003_35705___bin.27	232	2780	1312	1537
YuJ_2015___SZAXPI003424-12___bin.14	217	2739	1207	1577
YuJ_2015___SZAXPI015233-19___bin.67	208	2446	1027	1454
YuJ_2015___SZAXPI015252-43___bin.58	112	2592	1125	1523

Table 4: Number of contigs, CDS counts, Hypothetical proteins counts and known proteins counts per sample

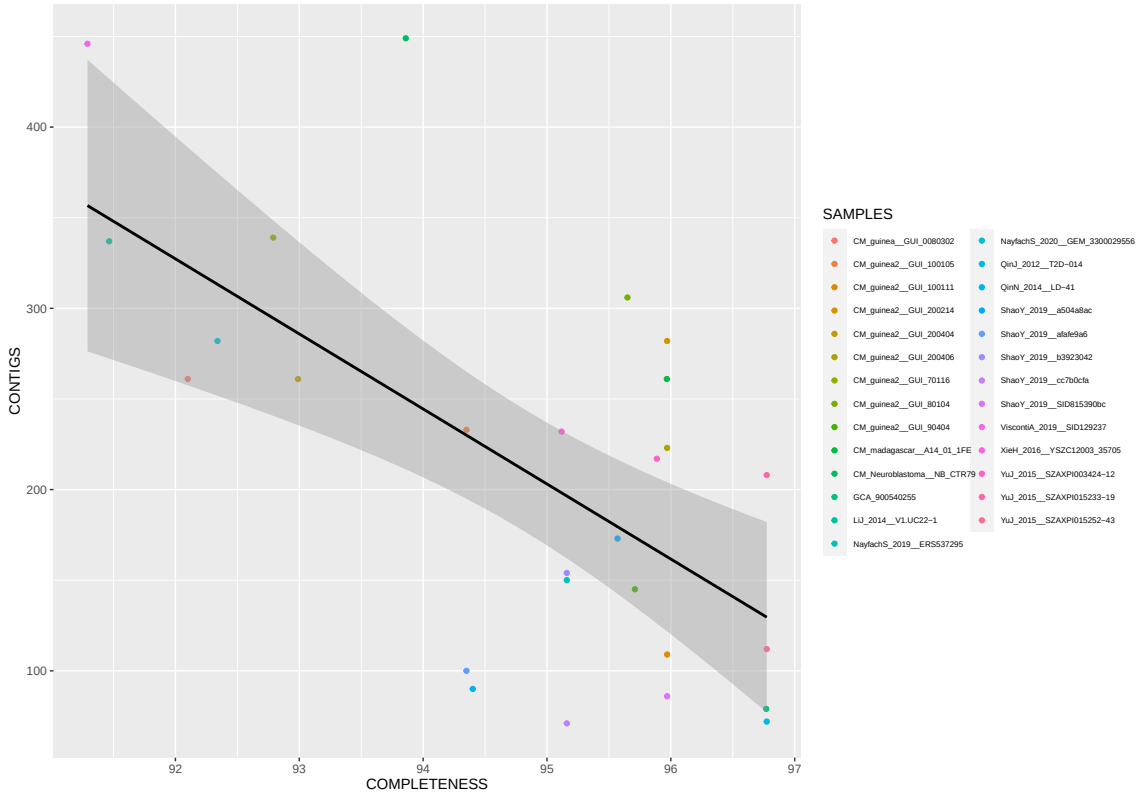

Figure 8: Contigs versus Completeness per genome (MAG or isolate)

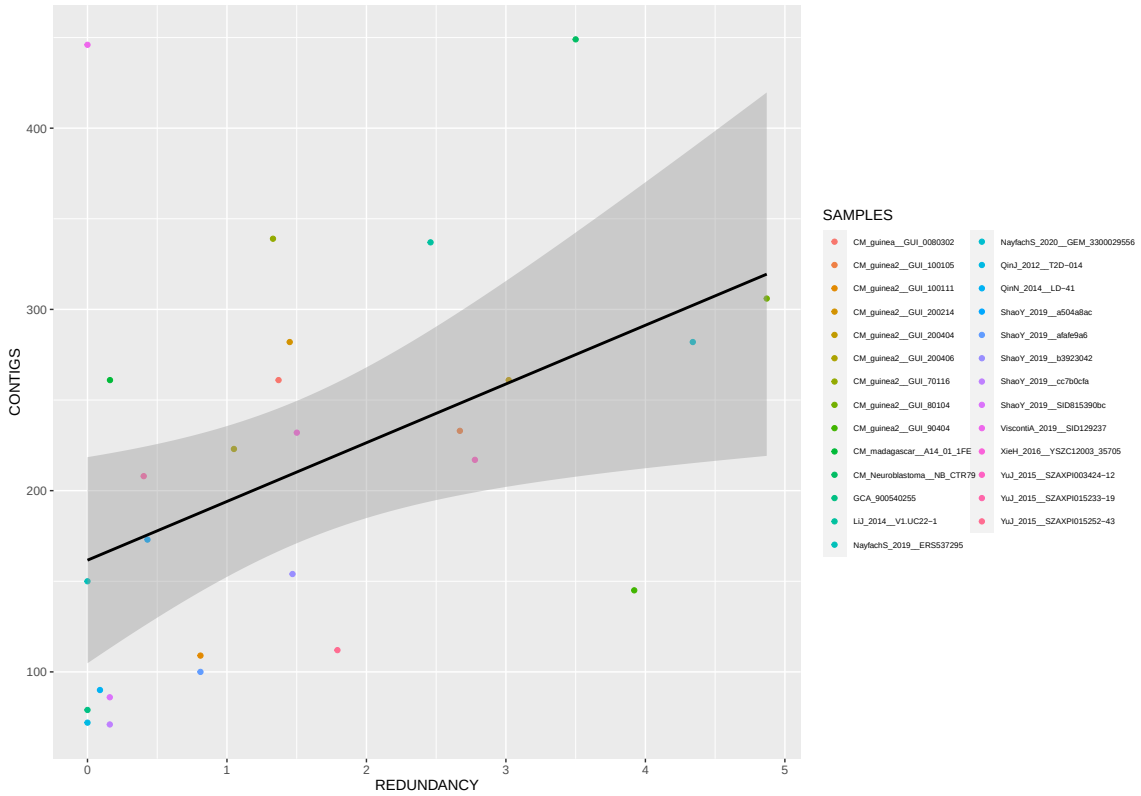

Figure 9: Contigs versus Redundancy per genome (MAG or isolate)

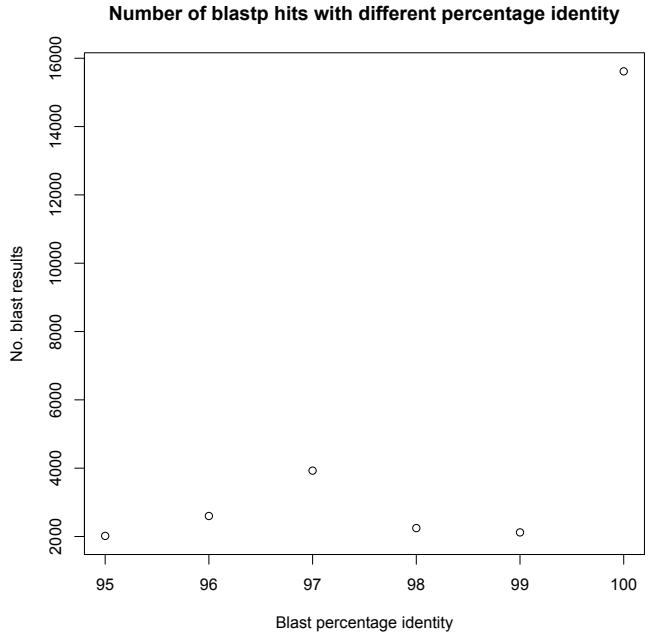

Figure 10: Number of Blastp hits with different percentage identity (using Blastp alignment)

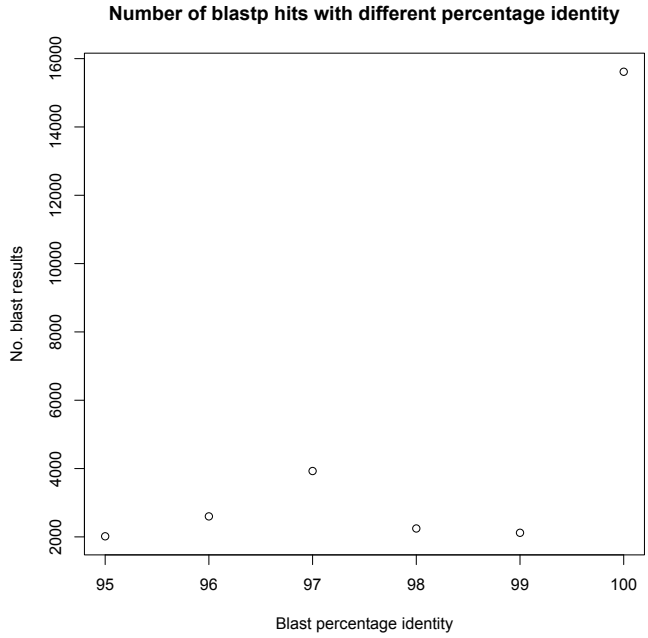

Figure 11: Number of Blastp hits with different percentage identity (using alignment with MAFFT and PRANK)

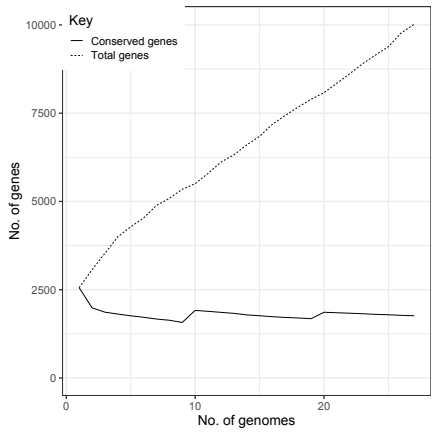

Figure 12: Conserved genes and total genes across pangenome: using alignment with MAFFT and PRANK

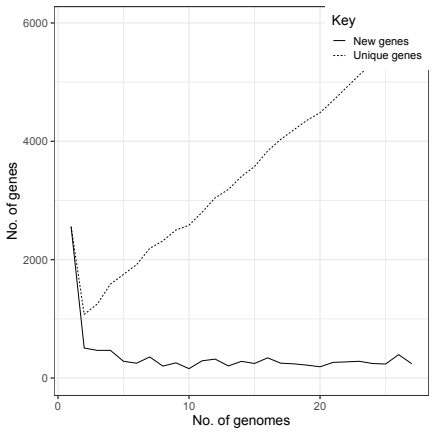

Figure 13: New genes and unique genes across pangenome (using alignment with MAFFT and PRANK)

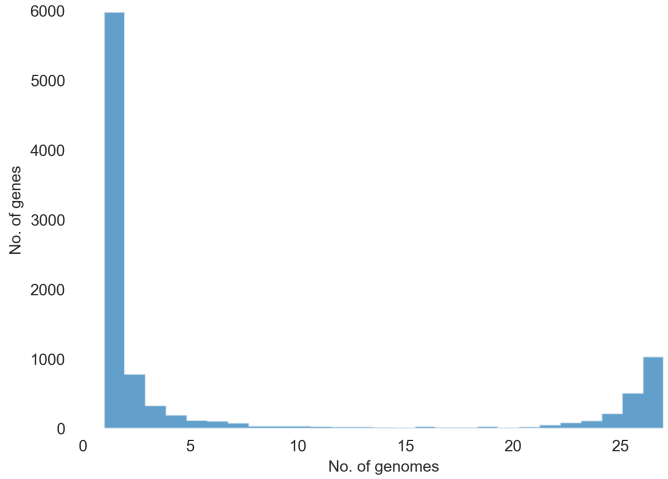

Figure 14: Frequency of genes across pangenome (using alignment with MAFFT and PRANK)

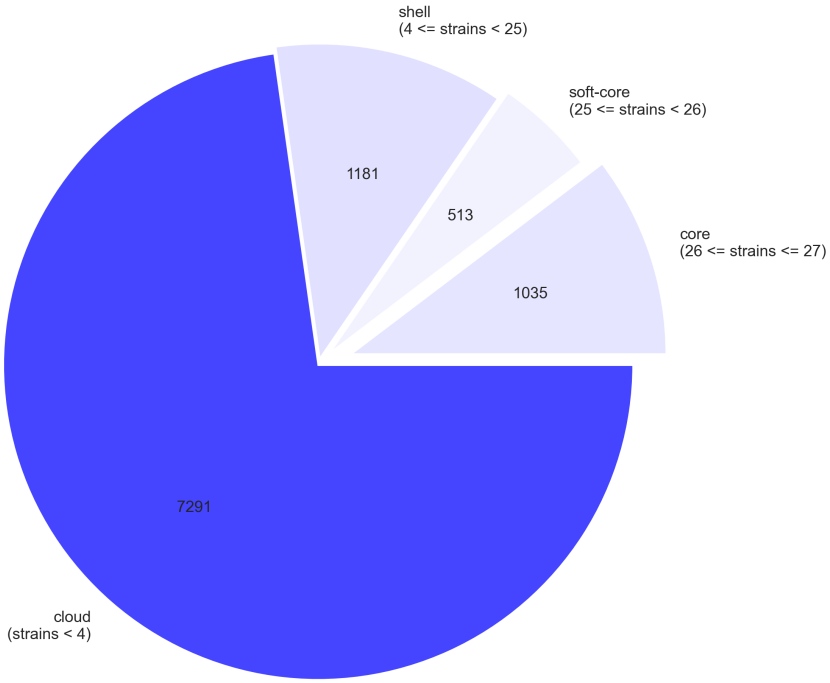

Figure 15: Pangenome genes composition: cloud (7291), shell (1181), soft-core (513) and core (1035) genes (using alignment with MAFFT and PRANK)

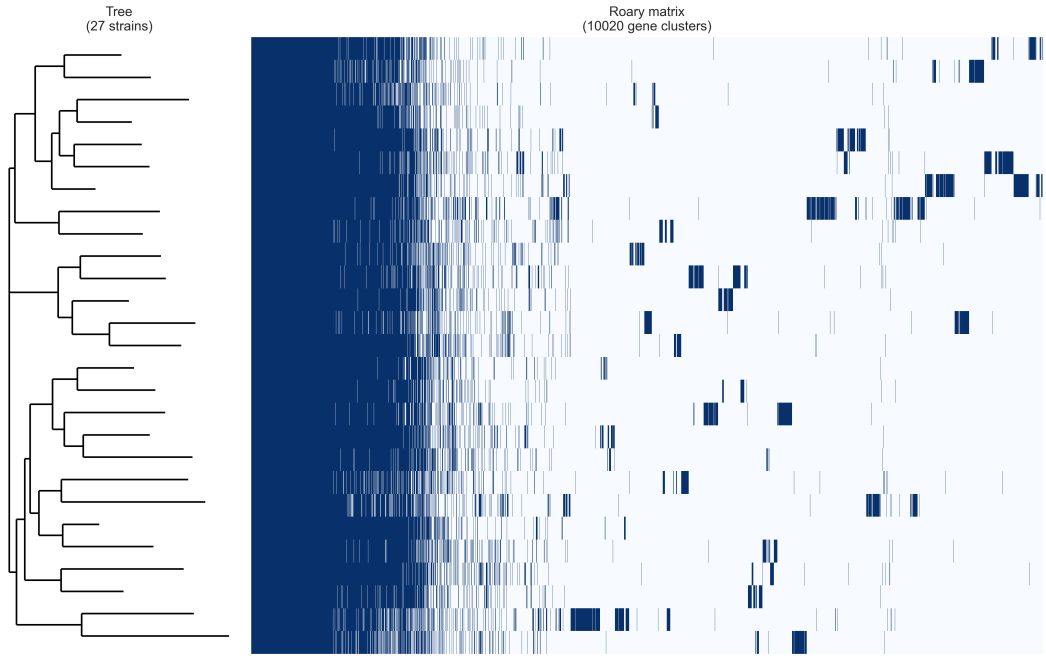

Figure 16: Heatmap of pangenome (using alignment with MAFFT and PRANK). Dark blue represents gene presence; light blue represents gene absence.The x-axis represents 10020 genes clusters and y-axis represents 27 strains in dendrogram. The gene clusters at left in dark blue depicts core genes.

#input_bin	uSGB taxonomical distance
CM_madagascar___A14_01_1FE___bin.13	uSGB_6179|t___SGB6179: 0.01834493125
LiJ_2014___V1.UC22-1___bin.24	uSGB_6179|t___SGB6179: 0.01812152649193548
QinJ_2012___T2D-014___bin.33	uSGB_6179|t___SGB6179: 0.015865366411290324
QinN_2014___LD-41___bin.25	uSGB_6179|t___SGB6179: 0.015955060927419357
XieH_2016___YSZC12003_35705___bin.27	uSGB_6179|t___SGB6179: 0.019111369879032256
YuJ_2015___SZAXPI003424-12___bin.14	uSGB_6179|t___SGB6179: 0.02015145318548387
YuJ_2015___SZAXPI015233-19___bin.67	uSGB_6179|t___SGB6179: 0.01643216536290323
YuJ_2015___SZAXPI015252-43___bin.58	uSGB_6179|t___SGB6179: 0.01823083766129032
CM_guinea2___GUI_100105___bin.75	uSGB_6179|t___SGB6179: 0.02398411612903226
CM_guinea2___GUI_100111___bin.34	uSGB_6179|t___SGB6179: 0.019066705887096774
CM_guinea2___GUI_200214___bin.86	uSGB_6179|t___SGB6179: 0.019749577943548386
CM_guinea2___GUI_200404___bin.54	uSGB_6179|t___SGB6179: 0.01768373052419355
CM_guinea2___GUI_200406___bin.2	uSGB_6179|t___SGB6179: 0.021495570564516127
CM_guinea2___GUI_70116___bin.90	uSGB_6179|t___SGB6179: 0.01877004794354839
CM_guinea2___GUI_80104___bin.57	uSGB_6179|t___SGB6179: 0.01979633435483871
CM_guinea2___GUI_90404___bin.43	uSGB_6179|t___SGB6179: 0.01611623120967742
CM_guinea___GUI_0080302___bin.8	uSGB_6179|t___SGB6179: 0.017641257741935482
CM_Neuroblastoma___NB_CTR79___bin.26	uSGB_6179|t___SGB6179: 0.01964306290322581
GCA_900540255	uSGB_6179|t___SGB6179: 0.017724237499999997
NayfachS_2019___ERS537295___bin.57	uSGB_6179|t___SGB6179: 0.019090847177419355
NayfachS_2020___GEM_3300029556___bin.6	uSGB_6179|t___SGB6179: 0.023584942983870965
ShaoY_2019___a504a8ac-7ae6-11e9-a106-68b59976a384___bin.8	uSGB_6179|t___SGB6179: 0.016583730443548387
ShaoY_2019___afafe9a6-7ae6-11e9-a106-68b59976a384___bin.6	uSGB_6179|t___SGB6179: 0.017236076975806452
ShaoY_2019___b3923042-7ae6-11e9-a106-68b59976a384___bin.23	uSGB_6179|t___SGB6179: 0.01742517806451613
ShaoY_2019___cc7b0cfa-7ae6-11e9-a106-68b59976a384___bin.21	uSGB_6179|t___SGB6179: 0.018078696693548384
ShaoY_2019___SID815390bc-7ae6-11e9-a106-68b59976a384___bin.19	uSGB_6179|t___SGB6179: 0.01674633721774194
ViscontiA_2019___SID129237___bin.45	uSGB_6179|t___SGB6179: 0.019415520564516127

Table 5: Taxonomical distances of MAGs and *Clostridium* isolate. Kingdom: Bacteria, Phylum: Firmicutes, Class: Clostridia, Order: Clostridiales, Family: Clostridiaceae, Genus: Clostridium, Species: Clostridium_SGB6179.