Analysis of metagenomics data

Genome annotation, phylogenetic characterization and pangenome analysis of a set of MAGs associated with host metadata.

Linda Cova and Enrico Frigoli

Github: https://github.com/enricofrigoli/cmg_project.git

Computational Microbial Genomics (Prof. Segata)
University of Trento
A.Y. 2021-2022

Abstract

Metagenome-assembled genomes are high-quality putative genomes that derive from the assembly of contigs obtained from metagenomic sequencing experiments. They encode a great deal of knowledge and, even if it is not always possible to link them to a known bacterial species, they can be useful to help characterize other unknown single-genome bins. For this project, a set of putative genomes was analyzed and characterized as *Adlercreutzia equolifaciens*. Firstly, the genomes were annotated and the resulting information was exploited to determine the composition of the core genome. Subsequently, thanks to an alignment of the core genes, some phylogenetic structures were built but no evident grouping was identified with a clear link to the host data.

Introduction

Large-scale metagenome assembly has the purpose of reconstructing bacterial genomes from shotgun sequencing experiments. In order to recompose entire genomes from reads that are only hundreds of nucleotides long, the sequences undergo a pipeline of assembling and quality control procedures that lead to the metagenome-assembled genomes (MAGs). MAGs are high quality single-genome bins, which are groups of contigs binned together because they seem to belong to the same genome. Bins are subject to strict quality controls to be considered genomes, and they should have high completeness (above 90%) and low redundancy (below 5%).

This project has as starting point a set of 31 high-quality genomes: 29 of them are MAGs and 2 of them are reference genomes. The MAGs were obtained from five different published works on the gut microbiome [1, 2, 3, 4, 5] and all of them were previously reunited in one species-level genome bin (SGB). After having checked that all the genomes belong to the same species, the project was aimed at extracting as much knowledge as possible from these datasets with the tools indicated in the `Methods` section. The genomes were firstly annotated to retrieve information about the proteins that can be translated from them, as well as the non-coding regions. Subsequently, a pangenome analysis was performed to identify the number of core genes (genes present in all strains) and to determine whether the bacterial species in analysis has an open or closed pangenome. The pangenome is the entire set of genes presented by all strains composing a species, and it can be considered open if its size increases indefinitely when new individual are added to the genomes bin, while it is closed if the number of genes composing it reaches a plateau when a great deal of genomes is analyzed. Thanks to this analysis, genes were classified in core, soft core, shell and cloud based on the percentage of prevalence in the set of MAGs. Moreover, a phylogenetic structure was determined based on the presence/absence of accessory genes in each genome. A second type of phylogenetic analysis was performed starting from the global alignment of the core genes in order to produce a more accurate phylogenetic tree. In the end, the trees were compared with the host data associated with each MAG with the purpose of identifying some potential clusters.

Methods

Taxonomic assignment The taxonomic characterization was performed with PhyloPhlAn v. 3.0.60 ([7]). In particular, the selected tool was PhyloPhlAn metagenomic: a tool that can assign genomes and MAGs to species-level genome bins by computing their distance with a database of reference genomes. The parameters chosen for the analysis

were: `-nproc 4` to set the number of cores to be used by the program, `-n 1` to set the number of best taxonomic assignments to show, `--database_update` to ensure the tool uses the latest database version and `-d cmg2122` to set the database of markers to be used. This analysis was performed on all the sequences of the set and the tab separated file obtained as output contained the complete taxonomy, from kingdom to species, of each MAG.

Genome annotation The genomes were annotated with Prokka software v. 1.13 ([8]). The option `--kingdom Bacteria` was selected to specify that the genome used ad input is bacterial, since Prokka also annotates archaeal and viral genomes. This procedure was applied to all the sequences of the set. As output, a series of files were obtained for each MAG and three of them were further analyzed in this project: the annotation file (.gff) was used for the pangenome analysis, the text and tab separated files were used to visualize some statistics with R (link to the script).

Pangenome Analysis The species' pangenome was obtained using Roary v.3.7.0 ([9]) taking Prokka annotation files (.gff) as input. The following parameters were given: `-i 95` to set the minimum percentage of identity in the blastp alignement to 95%, `-cd 90` to set the prevalence in % MAGs for a gene to be considered core, `-e` to produce the core gene alignement file using PRANK ([10]), `-p 8` to specify the number of threads. A core gene alignement using MAFFT (`-n` parameter) was ran to check the difference between the two alignement engines.

The number of genes and related classification were retrieved from the output file `summary_statistics.txt` and `gene_presence_absence.csv`; plots were generated using two plotting script (link and link).

Phylogenetic Structure The resulting core gene alignement file (.aln) obtained with Roary was manually checked with Jalview v.2.11.21 and then processed with FastTree v.2.1.10 ([11]) using `-nt` as parameter (nucleotide alignement), which infers approximately-maximum-likelihood phylogenetic trees using the Jukes-Cantor model for nucleotide evolution. The R package `ggtree` ([12, 13, 14]) was used to plot the tree with the selected metadata.

Results and Discussion

Description of the set of bins

Information about the sequences used for this project were retrieved from the `metadata` and the `bin_data` files that were provided with the data. `Metadata` provides an insight about the experiments from which our MAGs were obtained. This includes data about the hosts from whom the samples were taken: their health conditions, age, gender and country, as well as information about the sequences: number and length of the reads, number of bases and other information concerning the experiments. The two reference genomes are curated by the NCBI and are not reported in this file, and for the other datasets not all the features are always provided. The `bin_data` file reports completeness and redundancy of the MAGs, including the reference genomes.

The set of MAGs come from 9 different datasets, all of them working on stool samples. For what concerns the patients involved, three main study conditions can be identified:

colorectal cancer (CRC), adenoma and control. The control group is the most numerous group, composed by 19 samples. The CRC and the adenoma groups present respectively 7 and 3 samples. Patients in the control group are cancer-free but not all of them are healthy: 7 of them suffer from fatty liver, hypertension, type two diabetes or a combination of the three. Moreover, all cancer patients are reported to be associated with one or more of these diseases (Fig. 1 (A)). All hosts are aged 35 or older, in particular 9 of them are considered senior (older than 65 years old). The samples are almost equally distributed between female and male patients (12 females and 13 males). The hosts come from 8 different countries, all of them from Europe with Austria being the most represented nation (18 samples), except for one patient from Argentina and one from Guinea. Only one sample was obtained from a non-westernized host and its country of origin is Argentina (Fig 1 (B-E)). All the analyzed MAGs were obtained from sequencing experiments carried out with the IlluminaHiSeq technology and the minimum read length indicated for each dataset shows that quality control was successfully performed as only high-quality reads longer than 30 nucleotides were retained.

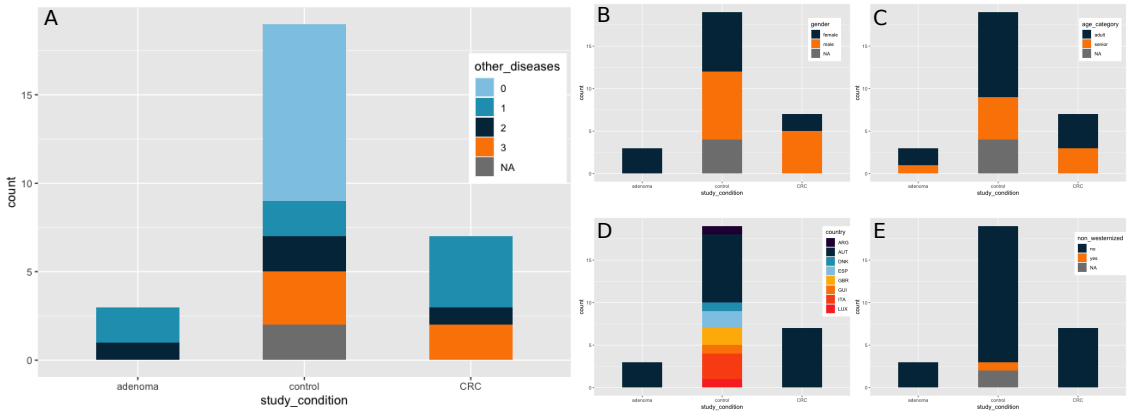


Figure 1: **Study conditions.** Number of samples per study condition (CRC, adenoma, control). Distribution of: A) other diseases, B) gender, C) age, D) country of origin and E) westernized/non westernized in each study condition group.

Completeness is the estimate of how completely a MAG represents a full genome based on the presence or absence of single-copy core genes, which are the genes found in the vast majority of genomes. Redundancy is the measure of how many single-copy core genes are found within a genome[15]. All MAGs analyzed have high completeness that spans from 90.33% to 100%. In addition to the reference genomes, there are 2 other datasets presenting 100% completeness. Moreover, all MAGs present low redundancy, from 0 to 4.48%, with 5 of them having zero redundancy reported (Fig. 2). These values indicate that each sequence has high quality and has been rightfully considered as a MAG. Redundancy can be interpreted as an indication for the presence of contamination, because high redundancy levels could mean that more than one population of bacteria were considered in the assembling of the genome. Despite the good redundancy values reported, it is not wise to completely exclude the possibility of contamination since redundancy is not the sole indicator of contamination.

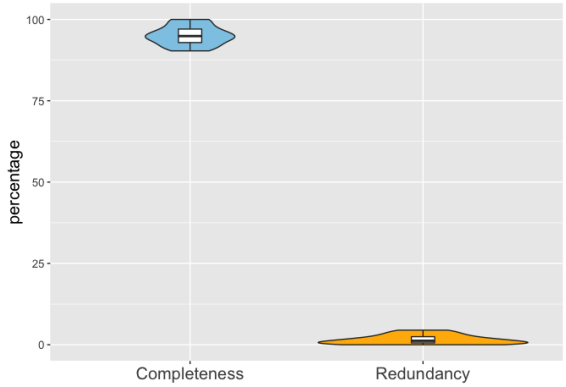


Figure 2: **Completeness and redundancy.** The violin plot shows how the percentages of completeness and redundancy are distributed among the set of MAGs. There are no genomes with completeness below 90% and no genomes with redundancy above 5%

Taxonomic assignment

The results of the PhyloPhlAn analysis conducted on all the MAGs considered for this project are reported in the table 1. The first analysis gave as result two MAGs assigned to a different Phylum (*Firmicutes*) with respect to all the others. These files were discarded from the project, since they were two low-quality incomplete genomes. Afterwards, the PhyloPhlAn output resulted consistent for each putative genome, confirming that they are all correctly clustered in the same species-level genome bin (SGB). The species of this SGB is *Adlercreutzia equolifaciens*.

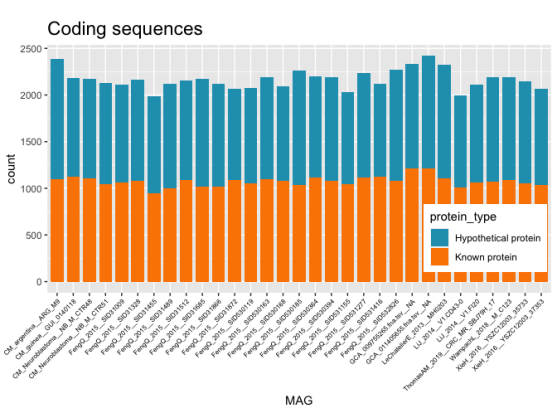
Kingdom	Bacteria
Phylum	Actinobacteria
Class	Coriobacteria
Order	Eggerthellales
Family	Eggerthellaceae
Genus	Adlercreutzia
Species	equolifaciens

Table 1: Taxonomic assignment of the set of MAGs

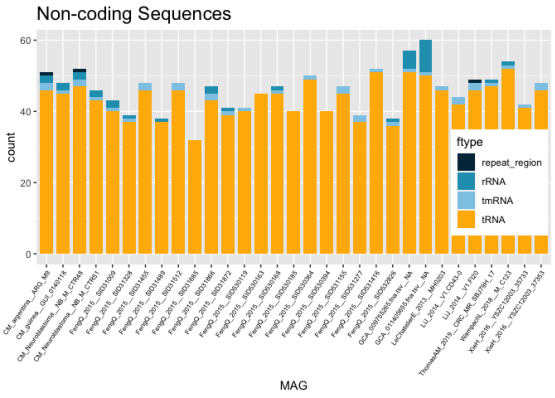
Adlercreutzia equolifaciens is a Gram positive, obligately anaerobic coccobacillus that can be found in human feces. It is capable of metabolizing equol from daidzein, a type of isoflavone found in soybeans and other similar plants[16]. Equol is a nonsteroidal estrogen produced by the gut microbiota of 30-50% of the human population. Some evidence suggests that it might play an important role in lipid metabolism[17].

Genome Annotation

Genome annotation was performed with *Prokka*. With this tool, the number of coding sequences (CDS), and non-coding sequences found within each genome were retrieved. CDS are divided in hypothetical and known proteins, while non-coding sequences are comprehensive of rRNAs, tRNAs, tmRNAs (bifunctional transfer-messenger RNAs) and repeat regions. Moreover, *Prokka* provides the gene symbol for each known protein and its sequence length.



(a) **Number of coding sequences.** The number of coding sequences annotated for each MAG, divided in hypothetical and known proteins.



(b) **Number of non-coding sequences.** The number of non-coding sequences annotated for each MAG, divided in repeat regions, rRNAs, tmRNAs and tRNAs.

Figure 3

The number of CDS in the set of MAGs goes to a minimum of 1988 to a maximum of 2544. For each genome, about half of the CDSs are non-characterized hypothetical proteins (figure 3a). For what concerns non-coding sequences, the vast majority of them is represented by tRNAs (figure 3b).

Pangenome Analysis

Pangenome analysis found 4670 total genes (figure 4c), among which 1017 were attributed to the *core* (above 90% prevalence in MAGs), 2074 to the *shell* (from 15% to 89% prevalence), and 1579 were classified as *cloud* (from 0% to 15% prevalence). The results were robust with respect to many rounds of computation.

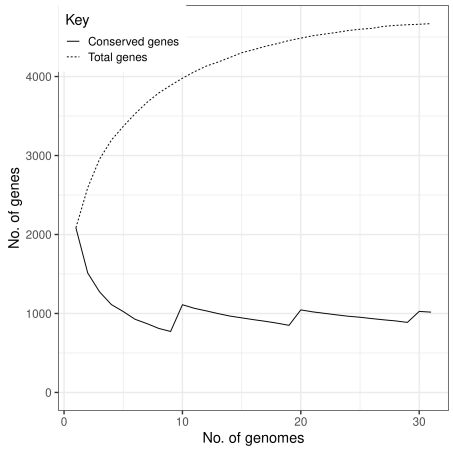
The number of conserved genes appears to reach a plateau (figure 4a) when the number of MAGs increases, suggesting that this species has a closed pangenome. This is further confermed by the trend of unique genes plotted against the number of genomes (figure 4b).

The plot of pangenome frequencies (figure 4d) shows the typical shape observed in microbiome samples. It is U shaped since there are conserved functions (core genes) that are present in every strain, and other genes that are very specific and are present in only one or few strains (unique genes), with very few genes in between. Both the pangenome matrix (figure 4e) and more detailed plots (figure 6), representing the number of genes plotted against the number of genomes considered, show an expected distribution.

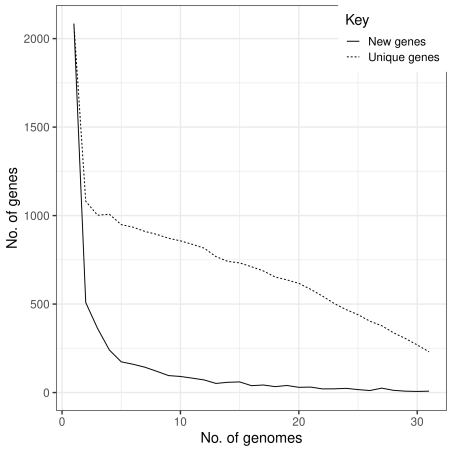
Phylogenetic Structure and association with host data

Significant differences in the topology were found between the tree obtained from the presence/absence of accessory genes and the one built from the core gene alignement (figure 5). Instead, MAFFT and PRANK produced core gene alignements that resulted in identical trees (figure 7).

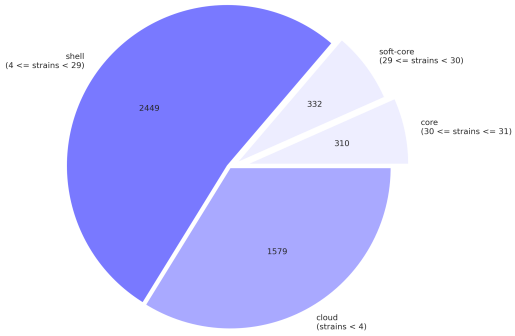
Three main clusters of strains were seen in the tree obtained from the alignement, but the low number of samples and the high fragmentation of information seen in the metadata impede the observation of finer clustering. For instance, nor clustering with respect to the country neither to the disease were visible.



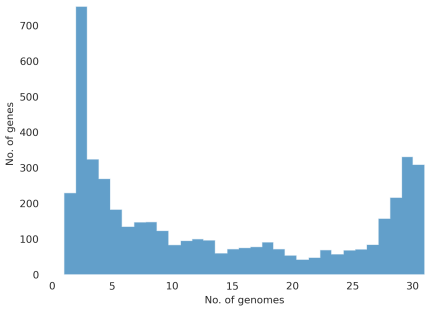
(a)



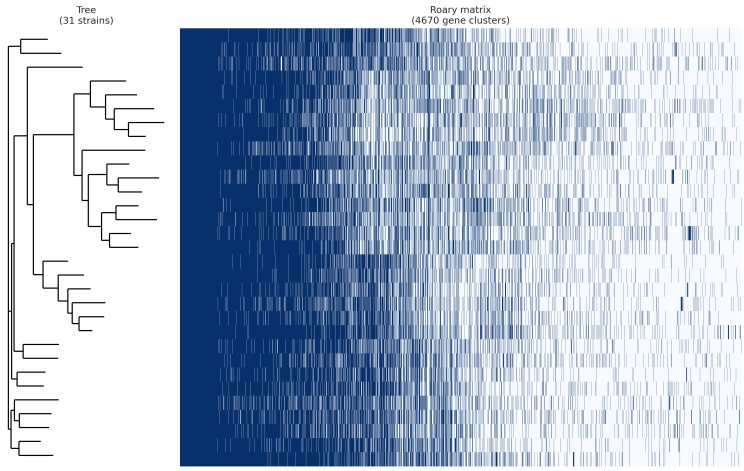
(b)



(c)

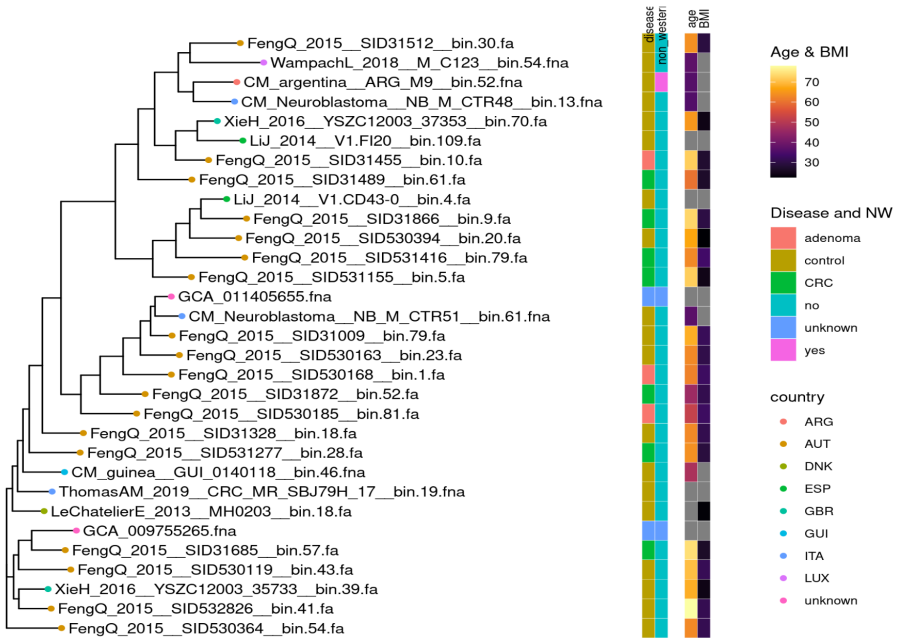


(d)

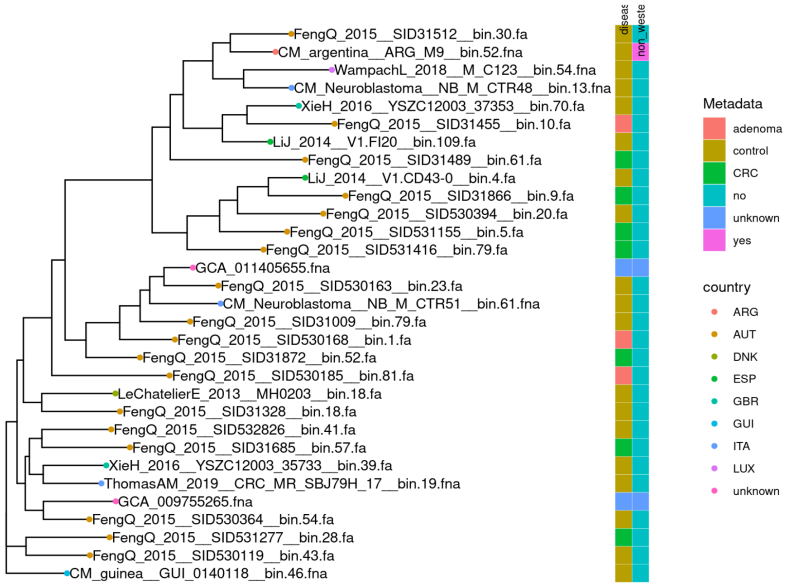


(e)

Figure 4: Collection of plots generated as output of the pangenome analysis.



(a)



(b)

Figure 5: Phylogenetic tree plotted with metadata, with the colour of the tip representing the country of provenence of the host. A significant difference in the topology is seen between the two trees, confirming the fact that the core gene alignement is a more reliable method to reconstruct the phylogeny. **(a)** Tree generated from the core gene alignement performed with PRANK; the heatmap represents (in order): the disease of the host, the classification as non-westernized (NW) the age and the BMI. **(b)** Tree generated from the presence/absence of accessory genes; the heatmap represents the country of provenence and the classification as non-westernized or westernized.

Conclusion

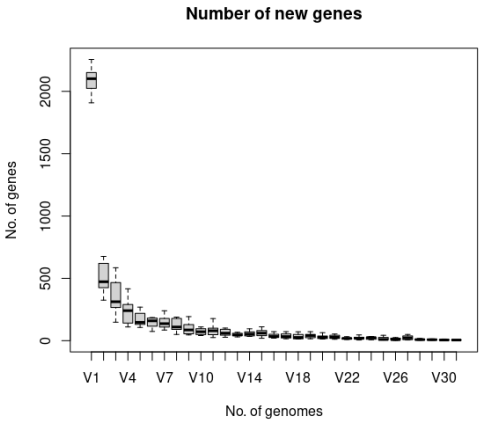
In this study, 31 high quality genomes (29 MAGs and 2 reference genomes) of the organism *Adlercreutzia equolifaciens* were characterized. The set of MAGs was derived from a pool of individuals that was heterogeneous with respect to country of provenence, health status, diseases, and age.

All the strains were correctly classified to the species. Genome annotation was performed to retrieve informations about the genomes, that were used to perform pangenome analysis, trough which the pangenome was found to be closed.

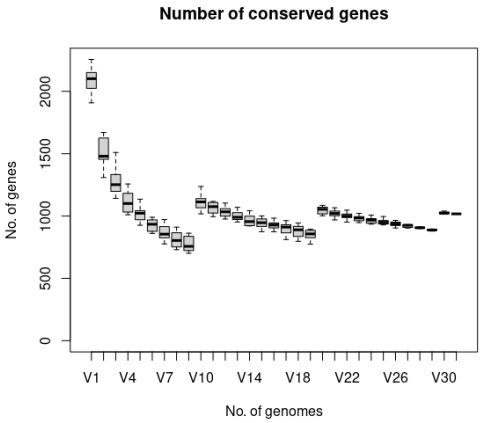
Subsequently, the phylogenetic structure was computed and, considering the host metadata, no significant clusters were found between the strains. As expected, the phylogeny retrieved from the presence or absence of accessory genes perfomed diffently with respect to the one obtained from the core gene alignement, with the last one to be considered more reliable. A bigger set of genomes will allow the detection of finer clustering with respect to host metadata.

Overall, our analysis showed consistent results with what was previously expected.

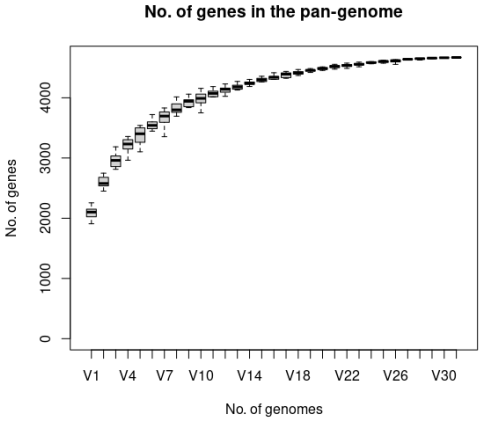
Supplementary data



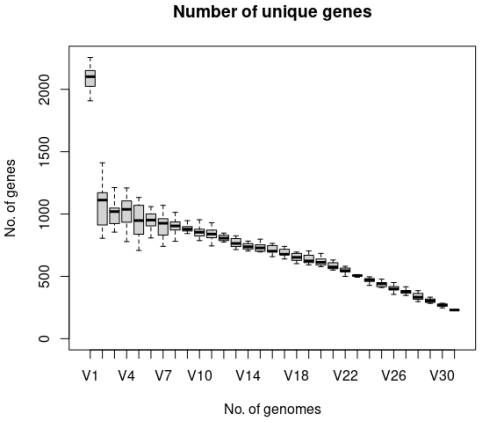
(a) Number of new genes



(b) Number of conserved genes



(c) Number of genes in the pan-genome



(d) Number of unique genes

Figure 6: Plots obtained from the output of the pangenome anotation with Roary. The number of new **(a)**, conserved **(b)**, total **(c)** and unique **(d)** genes show consistent and expected trend as the number of considered genomes increases.

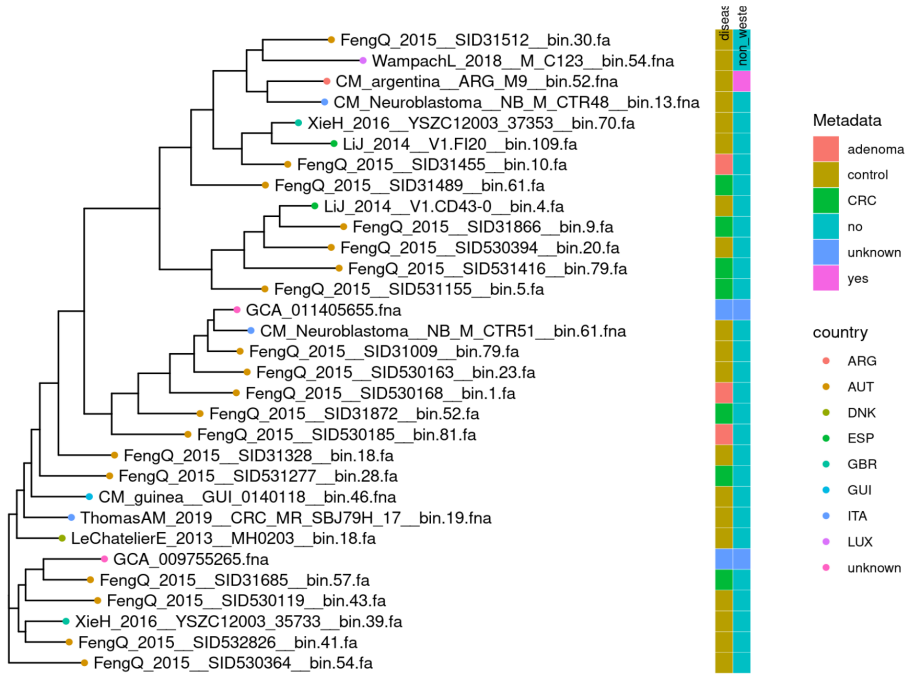


Figure 7: Tree generated from the core gene alignement obtained using MAFFT with the colour of the tip representing the country of provenence of the host; the heatmap shows the disease and the classification as non-westernized (NW) of the host.

References

[1] Q. Feng et al. "Gut microbiome development along the colorectal adenoma-carcinoma sequence". In: *Nat Commun* (2015).

[2] Emmanuelle et al. Le Chatelier. "Richness of human gut microbiome correlates with metabolic markers". In: *Nature* (2013).

[3] Cai X. et al. Li J. Jia H. "An integrated catalog of reference genes in the human gut microbiome". In: *Nat Biotechnol* (2014).

[4] Hailiang et al. Xie. "Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome". In: *Cell Syst* (2016).

[5] L. Wampach et al. "Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential". In: *Nat Commun* (2018).

[6] O. Tange. "GNU Parallel - The Command-Line Power Tool". In: *The USENIX Magazine* (2011). DOI: 10.5281/zenodo.16303. URL: http://www.gnu.org/s/parallel.

[7] F. Asnicar et al. "Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0". In: *Nat Commun* (2020).

[8] T. Seemann. "Prokka: rapid prokaryotic genome annotation". In: *Bioinformatics* (2014).

[9] Andrew J. Page et al. "Roary: rapid large-scale prokaryote pan genome analysis". In: *Bioinformatics* (2015).

[10] Ari Loytynoja. "Phylogeny-aware alignment with PRANK". In: *Multiple Sequence Alignment Methods*. 2014.

[11] Adam P. Arkin Morgan N. Price Paramvir S. Dehal. "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments". In: *PLOS ONE* (2010).

[12] Guangchuang Yu. "Using ggtree to Visualize Data on Tree-Like Structures". In: *Current Protocols in Bioinformatics* (2020). DOI: 10.1002/cpbi.96.

[13] Guangchuang Yu et al. "Two methods for mapping and visualizing associated data on phylogeny using ggtree." In: *Molecular Biology and Evolution* (2018). DOI: 10.1093/molbev/msy194.

[14] Guangchuang Yu et al. "ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data." In: *Methods in Ecology and Evolution* (2017). DOI: 10.1111/2041-210X.12628.

[15] Shaiber A. et al. Eren A.M. Kiefl E. "Community-led, integrated, reproducible multi-omics with anvio". In: *Nat Microbiol* (2021). DOI: https://doi.org/10.1038/s41564-020-00834-3.

[16] T. Maruo et al. "Adlercreutzia equolifaciens gen. nov., sp. nov., an equol-producing bacterium isolated from human faeces, and emended description of the genus Eggerthella". In: *Int J Syst Evol Microbiol* (2008).

[17] W. Zheng et al. "Compositional and functional differences in human gut microbiome with respect to equol production and its association with blood lipid level: a cross-sectional study". In: *Gut Pathog* (2019).