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

- PROJECT group SGB6179 -

Computational microbial genomics

Prof. Nicola Segata

Annalisa Xamin, Surya Hembrom, Surbhi Malhotra

University of Trento - Master in QCB - A.Y. 2021-2022

OUR SET OF BINS

- samples from SGB6179 (uSGB)
- 26 MAGs and 1 *Clostridium* uncultured isolate
- around 90-95% Genome completeness
- and <5% redundancy

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

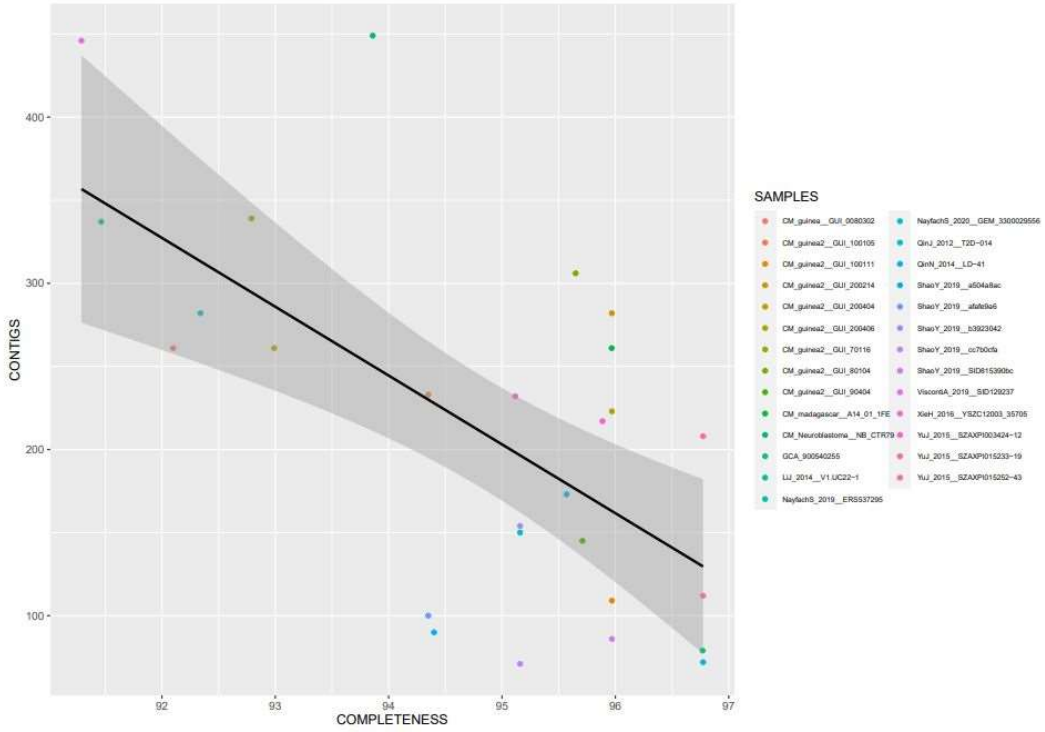


Figure 8: Contigs versus Completeness per genome (MAG or isolate)

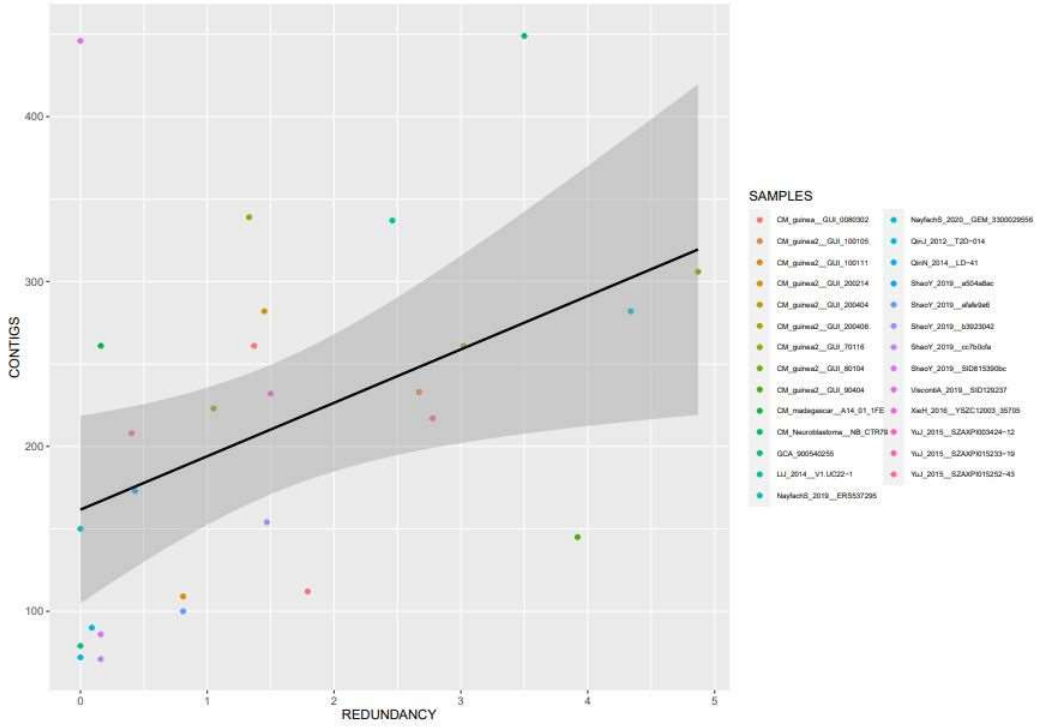


Figure 9: Contigs versus Redundancy per genome(MAG or isolate)

Genome annotation - Prokka

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

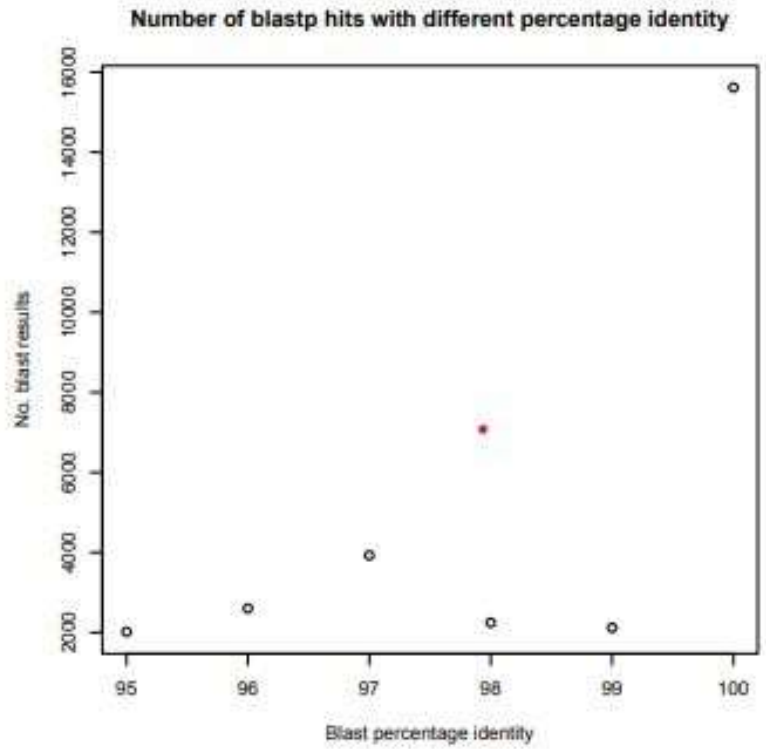


Figure 10: Number of Blastp hits with different percentage identity (using Blastp alignment)

PANGENOME ANALYSIS: ROARY

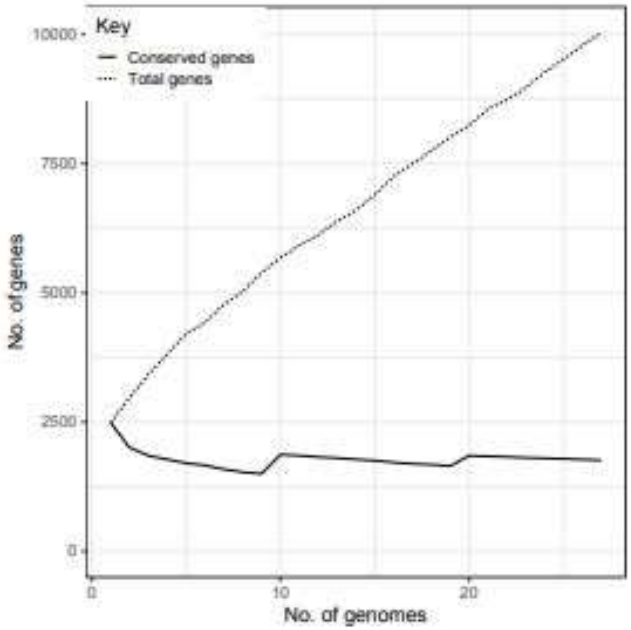


Figure 1: Conserved genes and total genes across pangenome

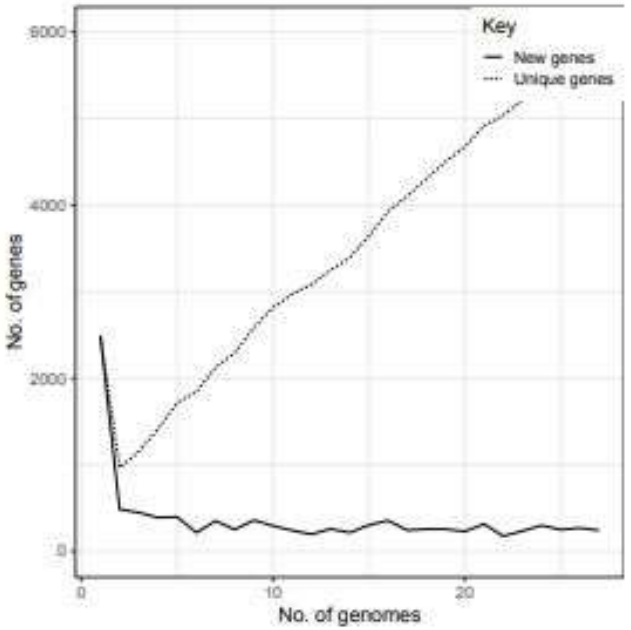


Figure 2: New genes and unique genes across pangenome

PANGENOME ANALYSIS - ROARY

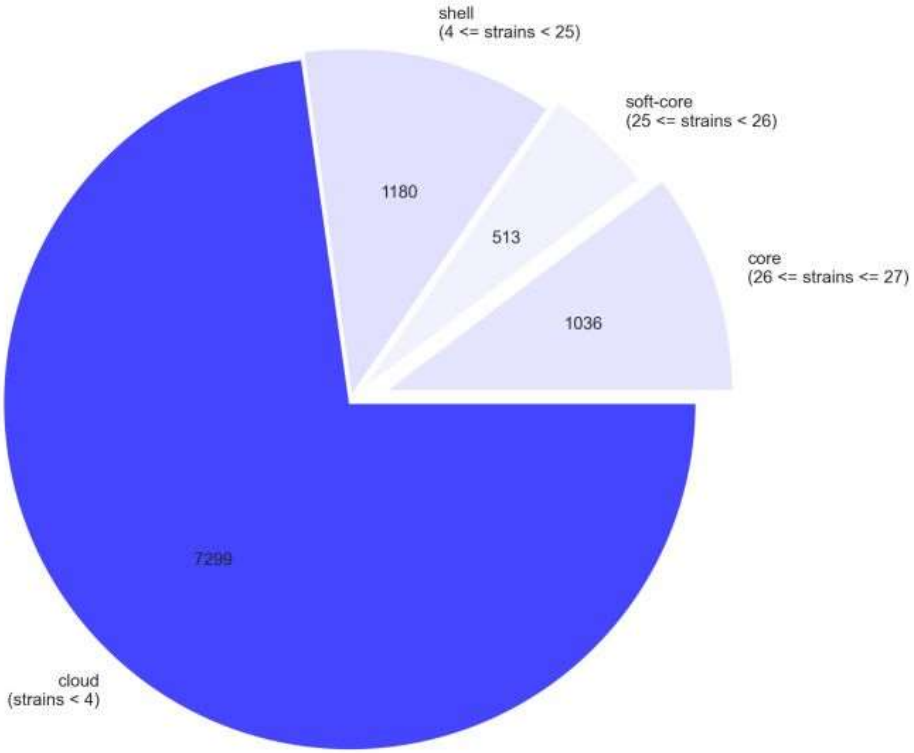
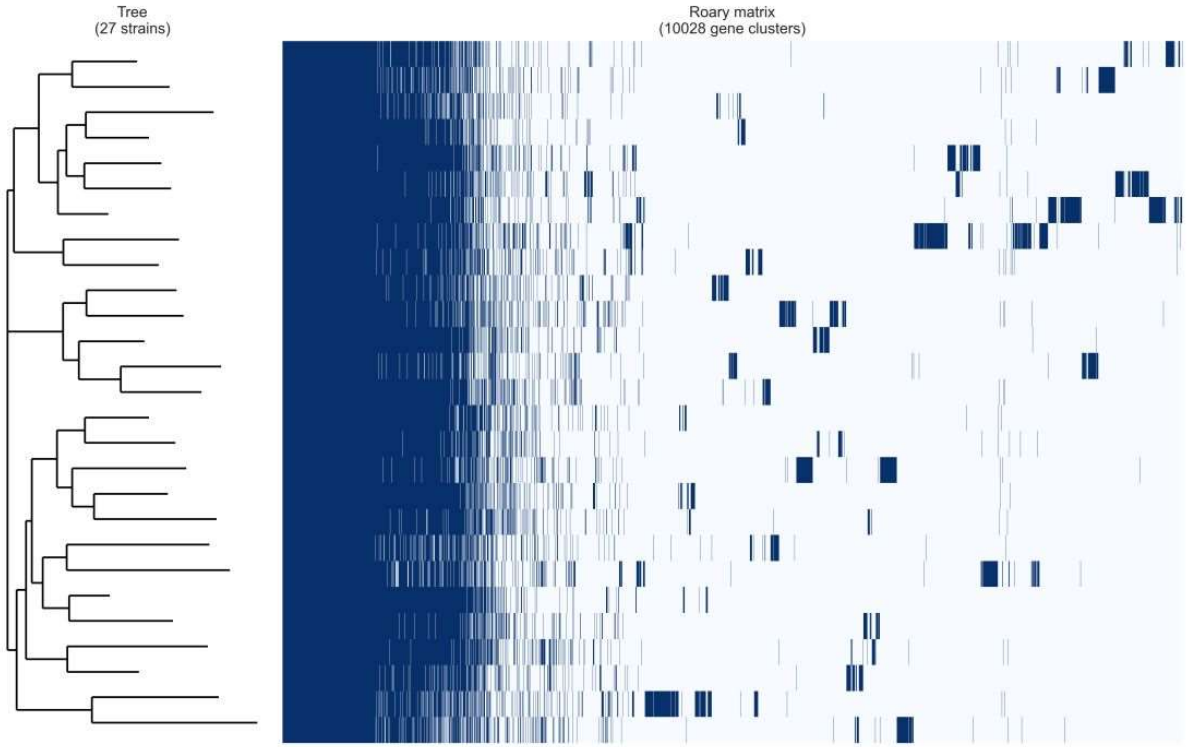


Figure 4: **Pangenome genes composition**: cloud (7299), shell (1180), soft-core (513) and core (1036) genes

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.



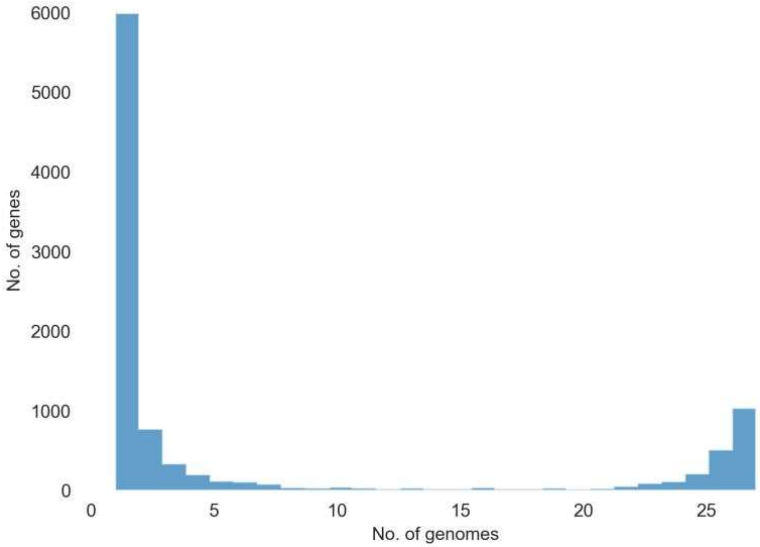


Figure 3: **Frequency** of genes across pangenome

PANGENOME ANALYSIS - ROARY

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

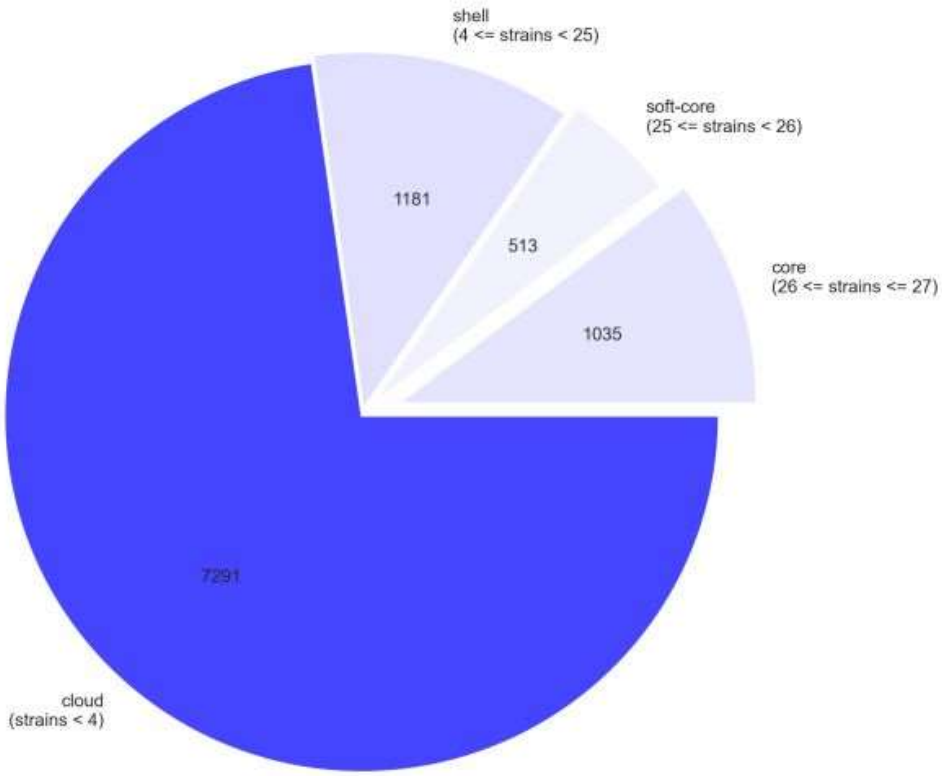
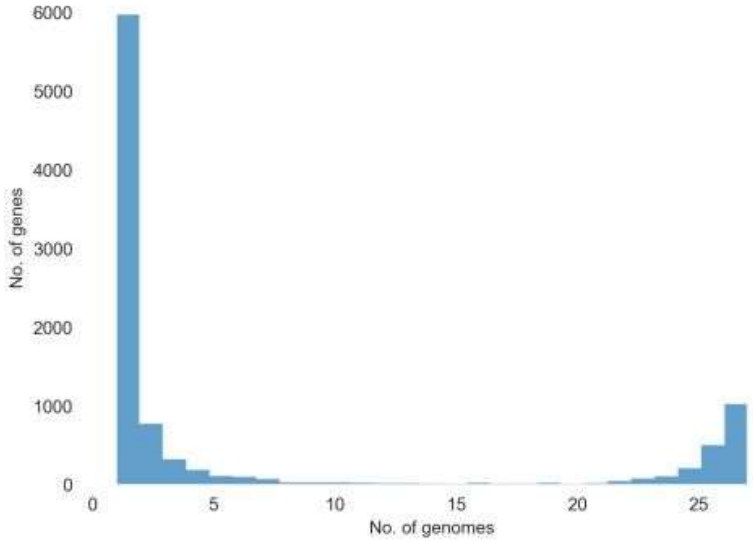


Figure 16: **Heatmap** of pangenome **(using alignment with MAFFT and PRANK).** Dark blue represents gene presence; light blue represents gene absence.The x-axis represents 10020 genes clusters and y-axis represents 27 strains in dendrogram. The gene clusters at left in dark blue depicts core genes.



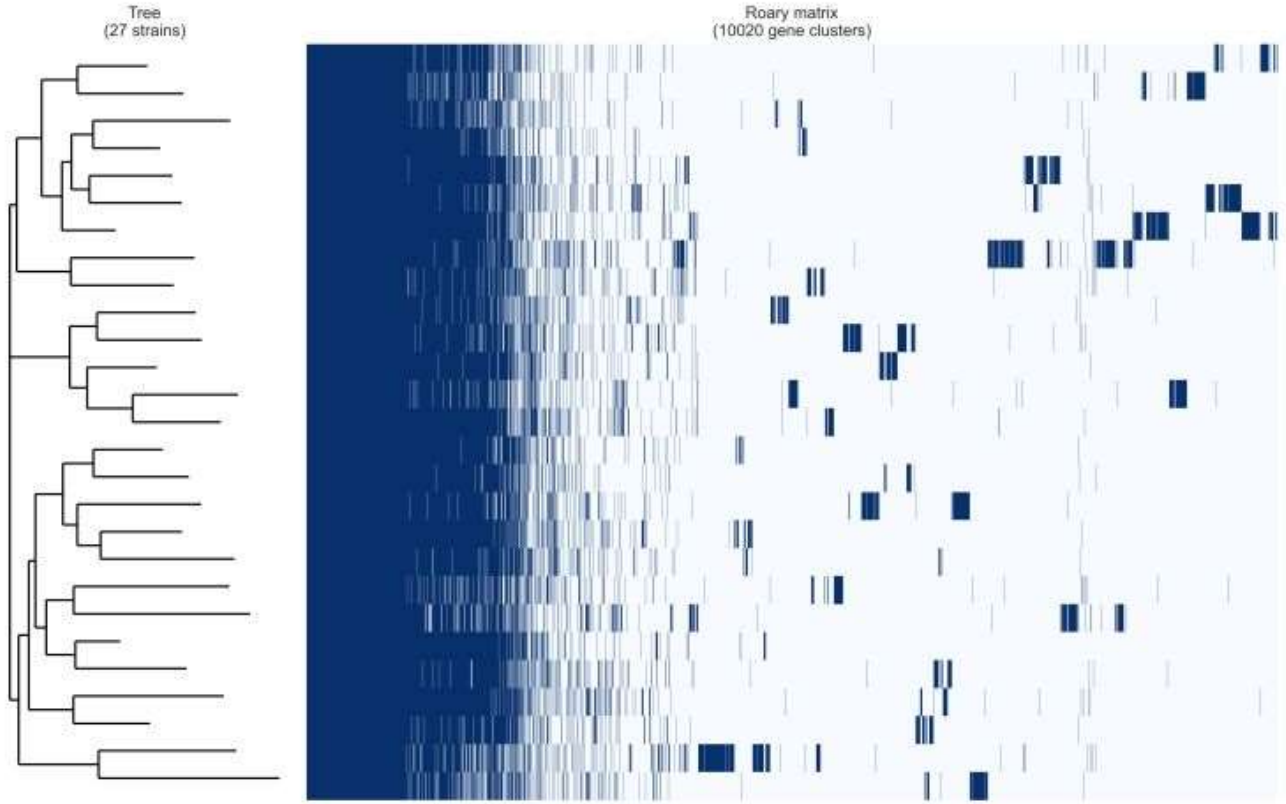


Figure 14: **Frequency** of genes across pangenome **(using alignment with MAFFT and PRANK)**

PHYLOGENETIC ANALYSIS - ROARY, FASTTREE &ITOL

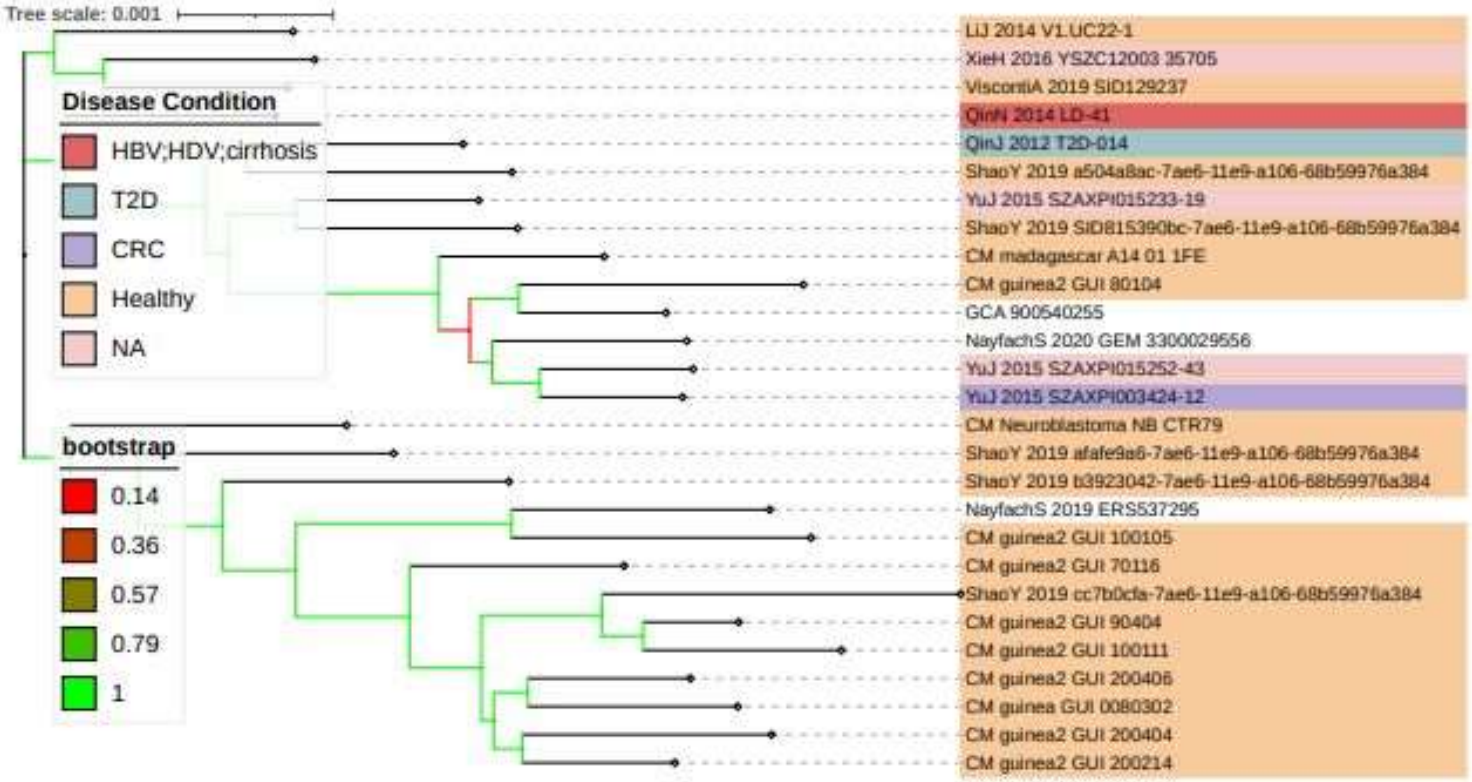


Figure 6: **Phylogenetic tree through core genes alignment.** The gut microbiome MAGs from healthy individuals are indicated in yellow, diseased patient with HBV (Hepatitis-V) + HDV (Hepatitis-D) + Liver cirrhosis in red, T2D (Type 2 Diabetes) in green and CRC (Colorectal Cancer) in purple. The Clostridium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

PHYLOGENETIC ANALYSIS - ROARY, FASTTREE & ITOL

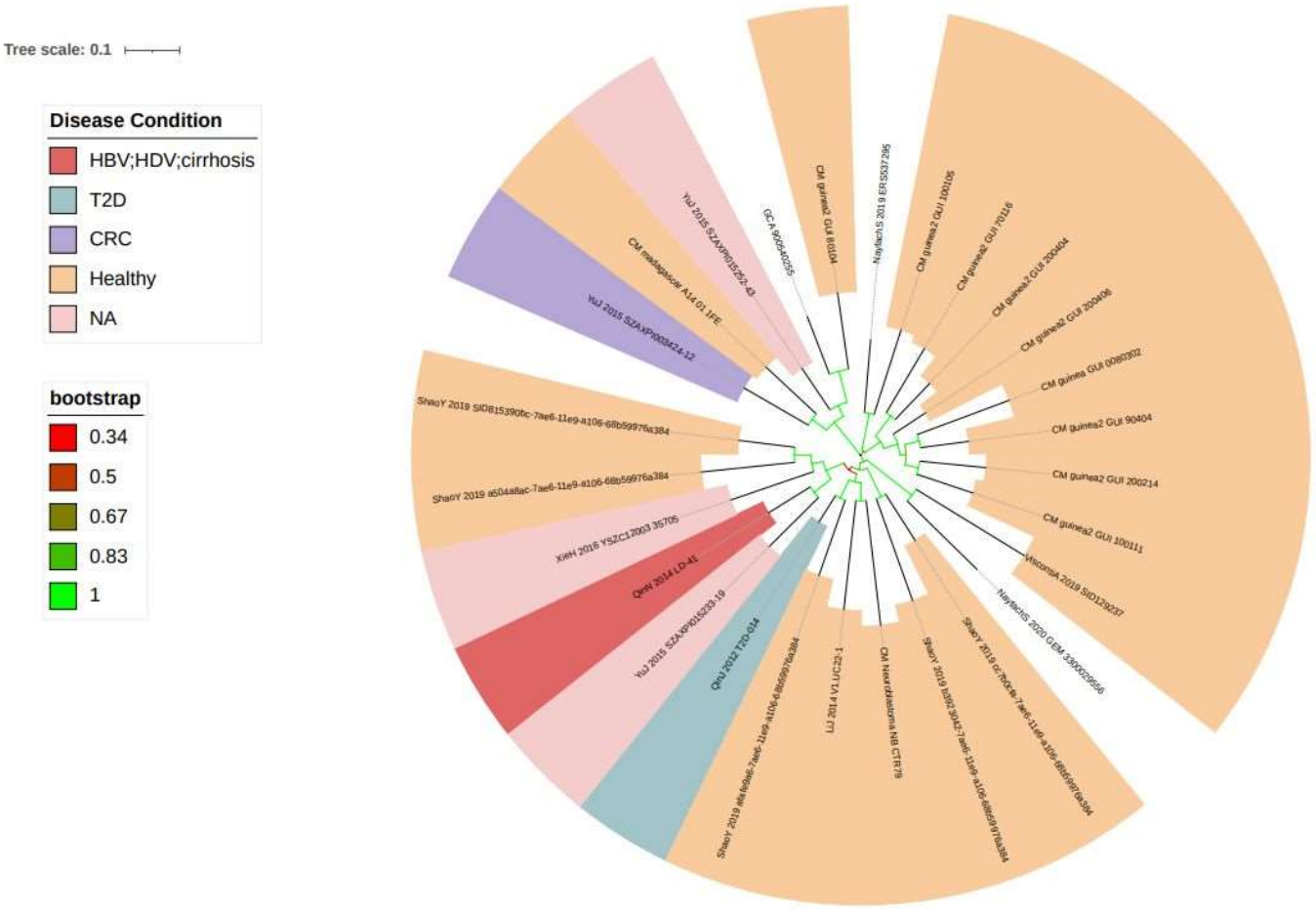


Figure 7: **Phylogenetic tree indicating presence/absence of accessory genes.** The gut microbiome MAGs from healthy individuals are indicated in yellow, diseased patient with HBV (Hepatitis-V) + HDV (Hepatitis-D) + Liver cirrhosis in red, T2D (Type 2 Diabetes) in green and CRC (Colorectal Cancer) in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

#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: 0.01772423749999997
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.