Norwegian University
of Life Sciences

Master's Thesis 2023 30 ECTS
Faculty of Chemistry, Biotechnology and Food Science

Human Gut Microbiome: GA-map Coverage and Genomic Categories in Humgut

Iben Amalie Lund Johansen
Chemistry and Biotechnology (M.Sc) - Bioinformatics

Abstract

Microorganisms in the gut are proven to affect health in multiple ways, and they have been studied extensively for the last decade. HumGut is a database containing microorganisms from the healthy human gut. The company Genetic Analysis (GA) has a GA-map dysbiosis test that aims to characterize microorganisms in the human gut to identify dysbiosis patients. This thesis provides an overview of the taxonomic and functional categorization of the human gut, gained by analyzing HumGut, and how well the GA-map spans the healthy human gut by investigating how it overlaps the database. Prodigal, Tax4FUN, and Diamond were used to acquire functional profiles for the HumGut genomes.

GA-map identifies microorganisms by matching GA-map probes to the 16S sequences in their genomes. Most genomes in HumGut do not have an identified 16S sequence. Still, the results show that the GA-map spans the majority of higher taxonomic ranks in HumGut. There are both taxonomic and functional parts missed by the map. The functional space missed by the GA-map seems to be relative similar to matched functional regions.

The genome categories in the HumGut collection are also evaluated in this thesis. The genomes in HumGut come from two sources: RefSeq and UHGG. Most of the UHGG genomes are Metagenome Assembled Genomes (MAGs). The findings show that MAGs may have lower quality on the 16S sequence than RefSeq-genomes, of which the latter is expected to have higher quality. The results also suggest that UHGG-genomes might have functional differences from other genomes.

Sammendrag

Det er bevist at mikroorganismer i tarmen påvirker helsen på flere måter, og det har blitt forsket mye på disse organismene det siste tiåret. HumGut er en database over mikroorganismer fra frisk human tarm. Bedriften Genetic Analysis (GA) har en GA-map dysbiose test med mål om å karakterisere mikroorganismer i human tarm for å identifisere dysbiose pasienter. Denne masteroppgaven presenterer en oversikt over human tarm gjennom taksonomisk og funksjonell kategorisering, oppnådd ved å analysere HumGut, og hvor godt GA-map spenner frisk human tarm ved å undersøke hvordan den identifiserer HumGut-genomene. Prodigal, Tax4FUN og Diamond ble brukt for å få funksjonelle profiler.

GA-map identifiserer mikroorganismer ved å matche med 16S-sekvensen til genomet. De fleste HumGut-genome har ikke en identifisert 16S-sekvens. Likevel viser resultatene at GA-map dekker over de fleste høyere taksonomiske nivåene i HumGut. Det er både taksonomiske og funksjonelle regioner som ikke dekkes. De funksjonelle regionene som ikke blir dekket av GA-kartet, ser ut til å være hovedsakelig like funksjonelle regioner som blir dekket.

En annen del av oppgaven er å evaluere genomkategoriene i HumGut. Genomene i HumGut kommer fra to kilder: RefSeq, som har validerte genomer, og UHGG. De fleste genomene i UHGG er Metagenom sammensatte genomer (MAGs). Funnene kan tyde på at disse genomene har lavere kvalitet på 16S sekvensene enn RefSeq-genomene, som er forventet å ha høyere kvalitet. Resultatene kan også indikere at UHGG-genomer kan ha funksjonelle forskjeller fra andre genomer.

Acknowledgments

This thesis is a part of a cooperation between Genetic Analysis (GA) and The Norwegian University of life sciences (NMBU) as a part of NMBU's master's program in Chemistry and Biotechnology at the Faculty of Chemistry, Biotechnology and Food Sciences (KBM) at NMBU the spring of 2023.

I would like to thank the main supervisor, Lars Snipen, for his engagement, fast responses, discussions, feedback, and support throughout this thesis. I would also like to thank Pranvera Hiseni, co-supervisor from GA, for good conversations, feedback, and guidance throughout the project. Additionally, thanks to the GA team for their support in this thesis.

Lastly, thanks to friends and family for all the good times, the motivation, and the help I have gotten from you through five years at NMBU.

Table of contents

List of Abbreviations

ANI Average sequence identity

BLAST Basic local alignment search tool

COGs Clusters of Orthologous Groups of proteins

GA Genetic Analysis (company)

GFF General Feature Format

GA-map GA-map dysbiosis test

GO Gene Ontology

IBD inflammatory bowel diseases

IBS Irritable bowel syndrome

KEGG Kyoto Encyclopedia of Genes and Genomes

KO KEGG ortholog

MAG Metagenome Assembled Genomes

NCBI National Center for Biotechnology Information

NMBU Norwegian University of Life Sciences

ORF Open reading frame

PCA Principal component analysis

PC Principal component

PICRUSt Phylogenetic Investigation of Communities by Reconstruction of Unobserved states

PLS Partial least squares

RDP Ribosomal Database project

RefSeq: NCBI Reference Sequence

UHGG The Unified Human Gastrointestinal Genome

UniRef UniProt Reference Clusters

1. Introduction

1.1 Human Gut

The human gut has been studied extensively for the last decade, especially the microbiome. The microbiome is the collection of microorganisms in the human gut (Turnbaugh et al., 2009). Parts of the microbiome are heritable, but it is shown that most are shaped by environmental factors (Rothschild et al., 2018). The microbiota affects the body in different ways, locally, like gut immunity (Althani et al., 2016), and in more distant organs, like regulating immunological defense against viral infection in the lungs (Ichinohe et al., 2011). It can also affect anxiety (Diaz Heijtz et al., 2011) and pain perception (Amaral et al., 2008). Microorganisms in the human gut have been proven to be directly associated with various diseases and disorders, such as Alzheimer's disease, Parkinson's disease, schizophrenia, and many more (Hou et al., 2022).

Gut bacterial dysbiosis is deviations in a healthy gut microbiome (Casén et al., 2015). Disorders such as Chron's disease, ulcerative colitis, inflammatory bowel diseases (IBD), irritable bowel syndrome (IBS), obesity, nonalcoholic steatohepatitis, and type I and type II diabetes are all examples of disorders associated with dysbiosis (Casén et al., 2015). Some of these disorders are quite prevalent. For instance, IBS affects around 11 % of the global population and gives a reduced health-related quality of life (Lovell et al., 2012; Akehurst et al., 2002).

Although dysbiosis has been researched extensively, it is unknown whether dysbiosis is an effect or a casual factor (Kim et al., 2023). Knowing more about the human gut may open opportunities related to diagnostics, treatment, or prevention of various diseases.

1.2 Databases covering microorganisms in human gut

Even if the microorganisms in the gut are proven to affect diseases and have been researched through the years, much is still unknown about them. Part of the problem is that only around one-third of them are found in the majority of healthy individuals (Lloyd-Price et al., 2016).

The major problem is that most of the microorganisms in the human gut are challenging to cultivate in the lab and, therefore, difficult to get isolates from (Lagkouvardos et al., 2017).

The introduction of short-read Metagenome Assembled Genomes (MAGs) has partially solved the latter problem. A deep read coverage of the human gut will contain the DNA of (nearly) all microorganisms living there (Sangwan, 2016). Depending on how long each read is, MAGs can be separated into short-read and long-read MAGs, where short-reads typically have a size around 150-250 basepairs (Maguire et al., 2020). MAGs can be derived from these read coverages by assembling reads and binning the results. Assembling reads generate contigs. Contigs are contiguous genomic fragments that are longer than the raw reads. Binning tries to discover patterns that can tell whether two contigs belong to the same genome. These patterns are used to separate contigs into bins, and the finished bins are MAGs (Maguire et al., 2020). Most MAGs lack a 16S sequence, and due to high similarity and high volumes of short-read data, assembling 16S is complex and may have poor quality (Yuan et al., 2015).

The Unified Human Gastrointestinal Genome (UHGG, https://www.ebi.ac.uk/metagenomics) database is a genome database containing MAGs from the human gut, and isolates (Almeida et al., 2021). National Center for Biotechnology Information has a database that only includes isolate, called NCBI Reference Sequence (RefSeq). RefSeq sequences are curated and modified to ensure quality (Pruitt et al., 2007).

HumGut is also a genome database containing genomes from UHGG and RefSeq collection. To make HumGut, over 5700 healthy human metagenomes were screened for the containment of over 490,000 publicly available microorganisms from UHGG and RefSeq. Over 381,000 genomes were found in the samples, and their prevalence score has been computed (Hiseni et al., 2021). The prevalence score is the fraction of metagenomes containing the genome, with some minimum Average Nucleotide Identity (ANI). ANI is a measure of similarity between two genome sequences (Yoon et al., 2017). These genomes were then clustered at 97.5% sequence identity, resulting in 30691 clusters. For each cluster, the one with the highest prevalence was chosen as the cluster representative. (Hiseni et al., 2021)

In addition to genome databases, there are also marker databases. One example is mBodyMap, a marker database for microbes in humans and their association with health (Jin et al., 2022). Another example is GMrepo v2, a human gut microbiome database focusing on disease markers. GMrepo v2 contains disease markers identified between two phenotypes, for instance, healthy versus disease (Dai et al., 2022). There are also general databases, like GreenGenes (DeSantis et al., 2006), SILVA (Quast et al., 2013), and Ribosomal Database Project (RDP) (Maidak et al., 1997) that contains all kinds of 16S sequences and not only those in the human gut.

1.3 16S marker

A marker that is frequently used is the 16S rRNA marker. 16S rRNA is a profiling phylogenetic marker gene. Such marker genes are often used in phylogenetic studies (Langille et al., 2013). The marker is conservated due to slow evolving and can therefore be used to classify microorganisms into taxonomical categories, and taxonomic relationships can be discovered from similarities in the marker (Yarza et al., 2008). 16S rRNA marker is used in projects like "Landscape of Gut Microbiome – Pan-India Exploration," which maps the Indian gut microbiome (Dubey et al., 2018). 16S rRNA contains no functional information (Langille et al., 2013)

1.4 Functional profiling

1.4.1 Gene prediction

Markers like 16S and genome databases can give insight into the human gut but do not contain functional information (Langille et al., 2013). Functional profiling, often based on the coding genes in the microorganism (Börnigen et al., 2013) may provide insight into how the microbiome affects health and how the lack or abundance of one microorganism influences the individual. The coding genes can be predicted if this information is not known.

Microbial gene prediction is the prediction of protein-coding genes in an organism. When predicting genes, software tends to be over-sensitive and produces faulty genes (Dimonaco et al., 2022). Some software that predicts genes are GLIMMER, which uses interpolated Markov models (Salzberg et al., 1998), GeneMark, a tool designed to improve finding gene

boundaries by using hidden Markov models (Lukashin et al., 1998), and Prodigal, which uses the fifth-order Markov model and aims to reduce false positives (Hyatt et al., 2010).

1.4.2 Protein sequence alignment

Protein sequence alignment is to compare (predicted) proteins to a database of known proteins (reference database) to find similarities and homologies between sequences. The Basic local alignment search tool (BLAST) is one tool that does this (Altschul et al., 1990). BLAST first breaks down the sequence into small fragments called K-mers. The K-mers are then compared to the reference database. If a K-mer matches the reference database, the hits are extended to generate a more precise alignment. BLAST often returns several alignments for each protein. (Almutairy & Torng, 2017). DIAMOND is a different protein sequence alignment tool that only can be used for protein, developed to be faster than BLAST (Buchfink et al., 2021).

1.4.3 Functional categories

One way to do functional profiling is to obtain functional categories and assign genes to them. Some functional databases are Gene Ontology (GO) which contains scientific functional profiles for different organisms, including microbial genomes (Harris et al., 2004), Clusters of Orthologous Groups of proteins (COGs) that contains sequenced genomes of prokaryotes and unicellular eukaryotes based on orthologous relations (Tatusov et al., 2001) and Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology, that links the gene to a functionally categorized gene catalog (Kanehisa et al., 2002).

Several tools can link genomes to functional categories. Two of these tools are Phylogenetic Investigation of Communities by Reconstruction of Unobserved states (PICRUSt) and Tax4FUN2, which links the 16S sequence to KEGG orthologs (Langille et al., 2013; Wemheuer et al., 2020). Thus, 16S sequences can be used to gain functional information, even if it does not contain it. Tax4FUN2 utilizes the high similarities in 16S rRNA between genomes to estimate the function of the microorganism. It links 16S to functional annotation of closely related genomes to assign functional profiles. (Wemheuer et al., 2020). The database UniProt Reference Clusters (UniRef) can also be used for this purpose. UniRef contains cluster protein in a database, and its functional search is based on similarity (Suzek et

al., 2007). Hence, UniRef can only be used for proteins that are in or have high similarity to a protein in the database.

PICRUSt and Tax4FUN2 both give a vector for each genome containing the functional profile. The vector shows which functional category the genome contains. This can be put together to show how many times each functional category was found in the genome. Different genomes can be compared to find similarities and differences in which functional categories were found and how many times they were found.

1.5 Genetic Analysis

Genetic Analysis (GA) is a company focusing on the human microbiome. One of their products is the GA-map Dysbiosis Test (GA-map), a product which uses 16S to characterize the microorganism in the human gut. The map consists of DNA probes that target bacteria in the human gut. To make this map, both healthy and IBS and IBD individuals were tested, and it was shown that the GA-map was able to identify and characterize dysbiosis patients based on these probes. (Casén et al, 2015).

The probes in the map are DNA that targets regions on the 16S. Using a probe instead of sequencing makes the analysis rapid and enables the testing of many fecal samples (Casén et al., 2015).

In a joint effort with Norwegian university of life sciences (NMBU), Genetic Analysis has also made HumGut, a genome database containing microorganisms in the human gut. To make HumGut, samples from healthy individuals from different regions of the world were screened for the containment of available genomes from RefSeq and UHGG. The goal of HumGut was to make one genome collection a universal reference that can be used for human gut microbiota. (Hiseni et al, 2021).

1.6 Aims of the study

The aim is to get an overview of the human gut in terms of taxonomic and functional categorization by analyzing HumGut and to see how well the GA-map spans the healthy human gut by investigating how it overlaps with the genomes in the HumGut genome collection. This knowledge is of some importance to the GA company when it comes to understanding how well their probes cover different taxonomic and functional groups from healthy human guts worldwide.

HumGut consists of genomes from various sources and is presumed to be of varying quality. Thus, an aim is also to evaluate the quality of the genomes in the HumGut collection. This aspect is of more general importance as it signifies the current progress in uncovering the genomic information related to the human gut microbiome.

The aims of the study is to

1. Get a taxonomic overview of HumGut and find out how the GA-map covers the taxonomic classifications.
2. Get a functional overview of HumGut and find out how the GA-map spans over the functional space of HumGut
3. Evaluate the quality of the genomes in HumGut

2. Methods

All data analysis has been done using RStudio 4.1.0 (R Development Core Team, 2010), and all figures have been made using the ggplot2 package (Wickham et al., 2016), in addition to the cowplot-package (Wilkerson, 2021) for multi-panel plots. The package tidyverse (Wickham et al., 2019) is used throughout the thesis.

2.1 HumGut

HumGut is a project aiming to be a genome collection used as a universal reference for human gut microbiota (Hiseni et al., 2021). HumGut is made by screening samples from healthy individuals for the containment of genomes from two different databases. The containment of these genomes found is the data used in this project.

HumGut is a joint effort between the company Genetic Analysis (GA) and the Norwegian university of life sciences (NMBU). GA also has a product called the GA-map Dysbiosis Test (GA-map). This map uses probes that match 16S sequences to characterize the microorganisms in the human gut. Some of the aims of this thesis involve finding out how this map covers HumGut.

2.1.1 Data retrieval

The data used in this thesis are two datasets from the HumGut project, made available from the company GA due to confidence surrounding the GA-map probes. The first dataset contains the genomes for the cluster representatives for 30691 HumGut clusters (one representative for each cluster). Clusters are small taxonomic orders with a high sequence identity. In this context, members of the same cluster have a 97.5 % average sequence identity (ANI). The rows in the data set are the genomes, and an overview of the columns in the cluster representative (HumGut) dataset can be seen in Table 2.1.

Table 2.1. The table shows some information for each of the columns in the HumGut cluster representatives data frame (Hiseni et al., 2021)

Column name	Description
Ncbi_tax_id	Taxonomy id from taxonomy database from NCBI (https://www.ncbi.nlm.nih.gov/taxonomy/)
Genome_id	Unique ID for the genome
Cluster025	The cluster the genome belongs to, with 97.5 % sequence identity
Cluster025_size	Numbers of genomes in the same cluster025
Cluster05	A broader cluster the genome belongs to, with 95 % sequence identity
Cluster05_size	Numbers of genomes in the same cluster05
Prevalence_score	The average occurrence in 3534 healthy human gut screens.
Metagenomes_present	Numbers of metagenome the genome was found in, using 95 % sequence identity as threshold
Genome_size	Number of basepairs in the genome
GC	GC-content in the genome
Completeness	Estimated completeness of the genome in percent
Contamination	Estimated contamination of the genome in percent
Genome_type	Type of RefSeq or UHGG (Complete Genome, Chromosome, Scaffold and Contig for the former and MAG and Isolate for the latter)
Source	RefSeq (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/) or UHGG (https://www.ebi.ac.uk/metagenomics/)

ftp_download	Address (ftp) the genome is downloaded from
HumGut_name	Unique HumGut name for each cluster025
HumGut_tax_id	Unique HumGut tax id for each cluster025
Gtdbtk_organism_name	GTDB-tk (https://gtdb.ecogenomic.org/) organism name for the genome
Gtdbtk_tax_id	Artificially created tax-ids for GTDB-tk
Gtdbtk_taxonomy	Full GTDB-tk taxonomy
Ncbi_organism_name	Organism name from the NBCI taxonomy database (https://www.ncbi.nlm.nih.gov/taxonomy/)
Ncbi_rank	Rank at the NCBI database
Path	Folder with the downloaded genome
Genome_file	Name of the FASTA file in an archive with the genomes

A dataset containing the subset of HumGut-clusters that has a 16S sequence previously extracted by the tool Barrnap (Seemann, 2013) was also retrieved and downloaded. This dataset will be referred to as HumGut16S. HumGut16S contains several members from each cluster. This dataset also contains test probes, the GA-map with names assigned from GA for this project, and how they match the 16S-sequences. A probe is considered to match the genome if the probe is entirely complementary to a stretch of the genome´s 16S-sequence. HumGut16S was reduced to only contain observations matching known forward and reverse primers. An overview of the columns in HumGut16S can be seen in Table 2.2.

Table 2.2. The table shows column information for HumGut16S (Hiseni et al., 2021). A semicolon implies that several columns are fitting the same description.

Column name	**Description**
Cluster025	The cluster the genome belongs to, with 97.5 % sequence identity
Genome_id	Unique ID for the genome
Source	RefSeq (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/)

	or UHGG (https://www.ebi.ac.uk/metagenomics/)
Genome_type	Type of RefSeq or UHGG (Complete Genome, Chromosome, Scaffold and Contig for the former and MAG and Isolate for the latter)
Prevalence_score	The average occurrence in 3534 healthy human gut screens.
Tax_id	Taxonomy id from taxonomy database from NCBI for RefSeq genomes (https://www.ncbi.nlm.nih.gov/taxonomy/) (and NA, not available) for UHGG
Sequence	The 16S sequence
FwPrimer; RvPrimer	Whether the genome has, or does not have, match to known forward/reverse primers.
IG0005; AG0703; AG1687; IG0133; AG1152; IG0060; IG0020; AG0815; AG0865; AG0638; IG0028; IG0012; IG0058; IG0023; AG1034; AG0930; IG0044; IG0053; AG0974; AG0651; AG0931; AG1099; AG0732; AG1225; AG1226; AG0377; IG0063; AG0608; AG0895; AG0416; AG1698; IG0197; AG0581; AG0620; AG0393; AG0396; AG0686; AG0863; AG0515; IG0079; AG1046; IG0020; AG0815; AG0865; AG0638; IG0028; IG0012; IG0058; IG0023; AG1034; AG0930; IG0044; IG0053; AG0974; AG0651; AG0931; AG1099; AG0732; AG1225; AG1226; AG0377; IG0063; AG0608; AG0895;	The different probes. The columns contain 0, meaning that the genome is not targeted by the probe, or 1, meaning the genome is targeted.

AG0416; AG1698; IG0197; AG0581;
AG0620; AG0393; AG0396; AG0686;
AG0863; AG0515; IG0079; AG1046;
AG0777; IG0081; AG1061; IG0314;
AG0912; AG1661; IG0011

HumGut16S was reduced to only contain observations that match to known forward and reverse primers.

2.1.2 Classification of genomes in HumGut

To be able to find patterns that arose because of similar classification, the genomes were classified into phylum, family, genera, and species. Phylum and family were classified using branch_retrieve in the *microclass* package (Vinje et al., 2016). Observations without phylum were discarded.

To use branch_retrieve, a node table containing all tax ids and how they are related and a name table containing all tax ids and their corresponding name was downloaded. Branch_retrieve uses these tables to search after tax id and returns tax_ids for the decided taxonomic levels. Six tax ids had expired and were replaced by new ones found by searching manually in the National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/taxonomy/). One rank at the time, branch_taxid2name in the *microclass* package was used to find rank names for the genomes. Branch_retrieve returned NA as tax_id for some ranks. To have genus and species for all observations, these ranks were extracted from the "NCBI-organism name" column in the HumGut-dataset.

2.2 Probes matching the genomes

One of the aims of this study is to find how well the GA-map spans the healthy human gut by investigating how it overlaps with the genomes in the HumGut genome collection. This means how the GA-map probes match overlaps HumGut16S. To investigate this, the cluster each observation belongs to is used. This means that the found results mirror the cluster level and not every single observation. In addition to analyzing which probes target a cluster, the

genome type was controlled to find deviations between the different genome types. This is part of evaluating the quality of the genomes in the HumGut-collection.

A cluster is considered matched by the probe if at least one observation within the cluster has a 16S sequence with a stretch that is entirely complementary to the probe, with a maximum of one mismatch. The number of matches for each cluster is the sum of probes matching at least one cluster member. This means that the clusters with zero targets have no matched cluster members. A cluster matched three times might be three observations belonging to the same cluster, each matched once by different probes, or one cluster member matched by three probes. Cluster defined as this will be referred to as a single HumGut genome.

2.2.1 Genome sources and categories

The National Center of Biotechnology Information (NCBI) has a public database called Reference Sequences (RefSeq), which contains nucleotide and protein sequences that are validated to confirm accuracy. RefSeq sequences are curated and modified. RefSeq contains a significant taxonomic diversity (Pruitt et al., 2007), but only some microorganisms in the human gut can be easily cultivated (Nayfach et al., 2019). One way to discover and characterize new microorganisms is to perform de novo assembly of shotgun metagenomic reads into contig sequences and sort them after sequence coverage. This enables the recovery of potential genomes, called metagenome-assembled genomes (MAGs), that are part of The Unified Human Gastrointestinal Genome (UHGG) database. In this thesis, all the MAGs are short-read MAGs. UHGG also contains isolates with varying completeness (Almeida et al., 2021).

The RefSeq genomes in HumGut have previously been shown to be significantly better quality than the UHGG genomes (Hiseni et al., 2022), and it was decided to control if the database source affected the matching rate of the clusters. RefSeq is divided into different genome categories depending on the degree of fulfillment. Contigs are reads put together to form large fractions of the chromosome. Scaffolds are unlocalized or unplaced contigs that have been connected with some gaps. A chromosome contains a sequence for one or more chromosomes. The chromosome may be completely sequenced with no gaps, or chromosome coining scaffolds or contigs with the gaps between. Chromosomes may contain unplaced or

unlocalized scaffolds. Complete genomes have no gaps, no unplaced or unlocalized scaffolds, and are expected to be completed. (https://www.ncbi.nlm.nih.gov/assembly/help/)

The genome´s mode genome category, which is the genome category occurring most frequently within the genome, and matching information as either matched or not matched were assembled to find any association between genome category and probe match.

2.2.2 Matches between genomes and GA-probes

For each genome, it was found how many probes that have a match. This was done both for all genomes in HumGut16S and for only genomes from the RefSeq source.

2.2.3 GA-map probes

Match information was also found on probes basis. For each of the probes, it was counted how many clusters, species, and genera it targets, using genomes from the RefSeq source only.

The company gave the probes names anonymously to keep their real identity hidden due to confidence surrounding the probes. These names do not show information about the probe. The probes were therefore named after the genomes they matched. The name shows which lowest taxonomic rank above 75 % of the genomes the probe matches belongs to, using RefSeq genomes only.

This means that if over 75 % of the RefSeq matches belong to the same species, the probe was named after the species. To show what taxonomic group the probe was named after, the probe name started with s_ for species, g_ for genus, f_ for family, or p_ for phylum. No probes are identical, but several got identical names. These clusters had the number of genomes targeted added to the probe names to clarify that no probes are equal.

2.3 Functional profiling

To investigate how well the GA-map spans the healthy human gut includes investigating how it spans over the functional space of HumGut. To do this, genes were predicted for all the

genomes in both data sets using Prodigal, and functional profiles were built using Tax4FUN, Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologs, and Diamond. The genomes were then clustered using K-means.

2.3.1 Building functional profiles

The coding genes in the microorganism are needed to do functional profiling with the desired tools. As this was not available in the dataset, it was obtained using the Prodigal software. Prodigal is a genome annotation software built to reduce false positives. Prodigal does not require parameters adjusted to the organism (Dimonaco et al., 2022) and was made to work for most organisms (Hyatt et al., 2010).

A subset of the HumGut table was filtered only to contain the cluster centroids for the genomes present in HumGut16S. The cluster centroid is the genome within the cluster that was most prevalent in the human guts sequenced to make HumGut. This subset was filtered only to contain the three columns genome_id, path, and genome_file and downloaded as a tab-separated text file.

2.3.1.1 Gene prediction

Prodigal was performed for the cluster representatives using a script that takes a tab-separated text file like that as input. The software first achieved the whole genome using a path and genome files. To predict genes, prodigal utilizes elements like start codon and ribosomal binding site in addition to the open reading frame (ORF) (Hyatt et al., 2010).

Prodigal results in General Feature-files (GFF) and FASTA-files, one for each cluster representative. These files were read in using readGFF and readFasta in RStudio. The GFF-file contains a score. The higher the score is, the more likely it is that the predicted gene is real, according to the Prodigal premises. To find a score limit, three random DNA sequences were made that contained no genes, and the prodigal results of these three genomes were read. Both random and genome DNA had low, mid, and high GC content. DNA with a high content of GC generally contains fewer stop- and start codons and more non-real ORFs. These ORFs are often chosen as the real ones instead of the actual ORF. This means that, in general,

software like Prodigal predicts genomes with lower GC contents more accurately. (Hyatt et al., 2010). The random and genome-DNA were used to make histograms to decide on a score.

The fasta and GFF-files were filtered only to contain genes with a score over the limit. The resulting fasta-files were used in DIAMOND.

2.3.1.2 Assigning genomes into KEGG orthologs

DIAMOND is a protein aligner that was developed to be a fast alignment on behalf of sensitivity compared to its alternatives. The sensitivity has also improved through improvements in the software (Buchfink et al., 2021). DIAMOND aligns the proteins to a database. For each comparison, DIAMOND computes the similarity between the protein and the closest match in the database. In this project, DIAMOND aligns the proteins from the fasta-files to Tax4FUN2. Tax4FUN2 is a database that assigns functional profiles to proteins using KEGG orthologs (KOs) (Wemheuer et al., 2020). Thus, functional profiles can be built based on the similarity between the query proteins and the proteins in Tax4FUN2, which have assigned functions.

The functional profile for each genome is a vector, showing which functional categories the genome holds. All the genomes' profiles were set together to form a matrix. The functional categories extracted were used as a factor, so the table gives 0 if the genome does not possess the category. The resulting table had one column for each genome and one row for each functional category. The number in each cell showed how many predicted genes in that genome that holds this category. A table showing whether the gene was found in the genome (0 for not found and 1 for found, presence-absence matrix) was also made, and this was used for further analysis.

2.3.2 Analyzing the functional categories

The resulting matrix contained one column for each functional profile and one row for each genome. Since each genome has many coordinates, they are in a high-dimensional space. To be able to see patterns, principal component analysis (PCA) was used. In PCA, as few components as possible are used to explain as much variance as possible. The principal components are sorted, so the first contains the most information. (Pearson, 1901) The first

two principal components were used to find patterns in the data regarding how HumGut16S is distributed in HumGut, whether different phyla can be separated using functional profiles, whether genomes matched by and not matched by a phylum are functionally different, and whether genomes from the two different sources are functionally different.

The latter was also done using partial least squares (PLS), which aims to maximize covariance between the decided response and other variables (Wold et al., 2001). Contrary to PCA, PLS is supervised (Ruiz-Perez et al., 2020). This means that PLS relies on data with known responses to carry out its analysis.

PLS was performed using *plsr* in the pls package in RStudio (Mevik et al., 2021). The genome source is already known and was used as both training and test dataset. *Plsr* uses leave-one-out cross-validation as default, which was also used in this project. In this type of cross-validation, each observation is left out once for validation and used as training in the rest of the repetitions (Gourvénec et al., 2003).

To find what separates groups from each other, correlation was used. In RStudio, correlation by default is calculated with Pearson distance. Pearson correlation is a measure of linear correlation between two variables, in this context, between the source and the KOs, calculated one by one. Pearson correlation distance is always between -1 and 1. A higher absolute value means stronger association between the predictors and responders, in this case, between the KO and source (Schober et al., 2018).

2.3.3 K-means

To take more of the variance into consideration, the functional profile matrix was clustered with K-means. K-means pick one representative for each cluster. For each observation, it calculates the distance between the observation and each representative and assigns the observation to the closest representative. To get more accurate results, it starts over again with new representatives and aims to find the optimal clusters. In K-means clustering, the optimal clusters have a low within-cluster sum of squares. This means that the genomes in the K-means clusters are as similar to each other and as distinct from other clusters as possible (Hartigan & Wong, 1979).

3. Results

3.1 HumGut

HumGut is a genome database covering microorganisms in the healthy human gut. Some of the microorganisms HumGut contains are present in a subset called HumGut16S. This subset contains the microorganisms with 16S-sequence that the tool Barrnap (Seeman, 2013) was able to extract when HumGut was made. GA-map Dysbiosis test (GA-map) from the company Genetic Analysis (GA) consists of probes that match some of the genomes in HumGut16S. One of the aims of this study is to get a taxonomic overview of HumGut and find out how the GA-map covers the taxonomic span. Since GA-map probes match 16S-sequences, they can only match genomes in HumGut16S. Thus, to say something about how the map covers the taxonomic span of HumGut, it is necessary to know whether HumGut16S is representative of HumGut.

The HumGut-collection consists of 30536 clusters classified into 16 phyla, 346 genera, and 1025 species. These clusters are cluster representatives and will be referred to as genomes. Of these, 15 phyla, 300 genera, 904 species, and 4253 clusters are represented in HumGut16S. The phylum *Candidatus Thermaplasmatota* has no genomes in HumGut16S. Figure 1 shows how HumGut16S is distributed across the taxonomic rank of the phylum.

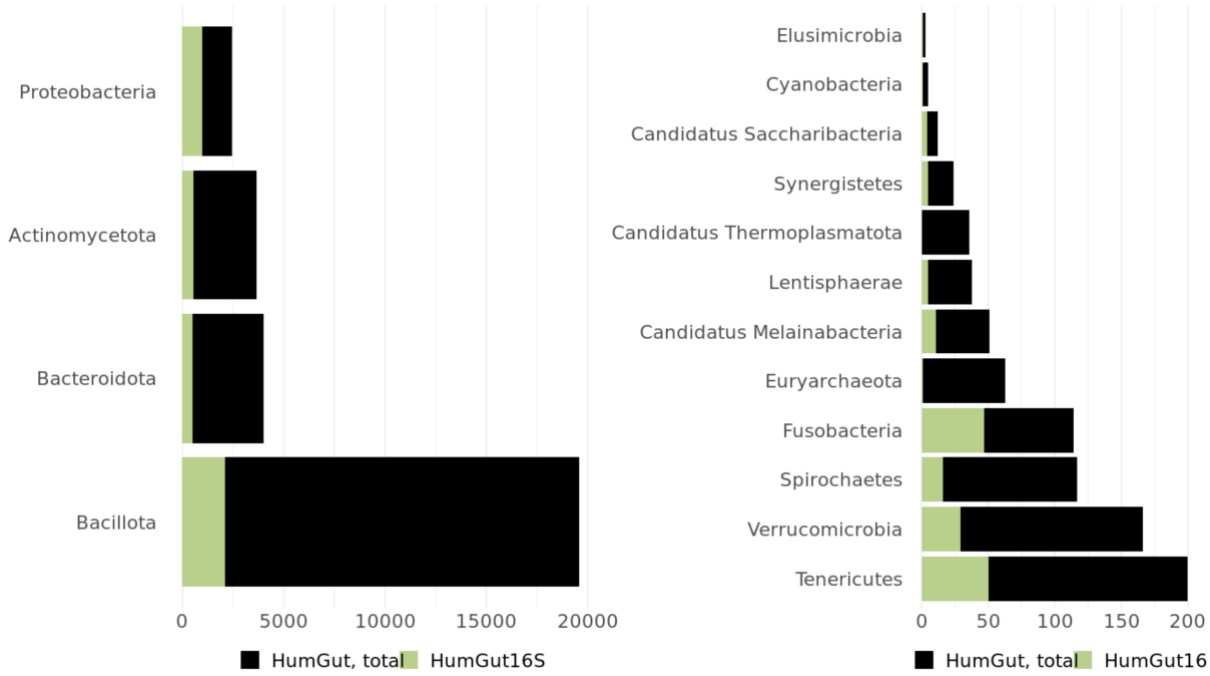

Figure 1. The figure shows the number of genomes of each phylum that is in HumGut16S for the four most frequent phyla (left panel) and the twelve other phyla (right panel) present in HumGut. The figure is split into two panels to have different scales along the x-axis. The bar for each phylum shows the total of clusters in the phylum in HumGut, and the green part is HumGut16S. One phylum, Candidatus Thermaplasmatota, has no genomes in HumGut16S.

3.2 How the GA-map overlaps HumGut

The GA-map contains probes that match 16S sequences in HumGut16S. A probe is considered to match the genome if the probe is entirely complementary to a stretch of the genome´s 16S sequence. One single mismatch is allowed because this does not stop their hybridization.

Matches are analyzed at the cluster level. This means that the number of matches on a genome is how many probes match at least one cluster member.

3.2.1 Genome categories

HumGut16S contains genomes assembled from two different sources, the National Center for Biotechnology Information (NCBI) Reference Sequences (RefSeq) and The Unified Human

Gastrointestinal Genome (UHGG). The two sources contain a total of six different genome categories in HumGut. One of these is metagenome-assembled genomes (MAGs) in UHGG. Assembling 16S-sequences in MAGs is challenging, which may lead to poor quality. UHGG also contains cultivated isolates with varying completeness. RefSeq has four different genome categories with different levels of completion, from contig (sequence contigs) and scaffold (connected contigs) to chromosome (sequence for one or more chromosomes) and completed genomes. RefSeq genomes are curated by NCBI and are previously shown to be significantly better quality than MAGs (Hiseni et al., 2022). One of this thesis aims is to evaluate the quality of the genomes.

The histogram in Figure 2 shows the fraction of not matched and matched genomes for each mode genome category, which is the genome category with the highest frequency within the genome. Genome categories from the UHGG source are colored in red (MAGs are darker than isolates), while genome types from RefSeq are grey. Lighter grey color indicates a more complete RefSeq genome.

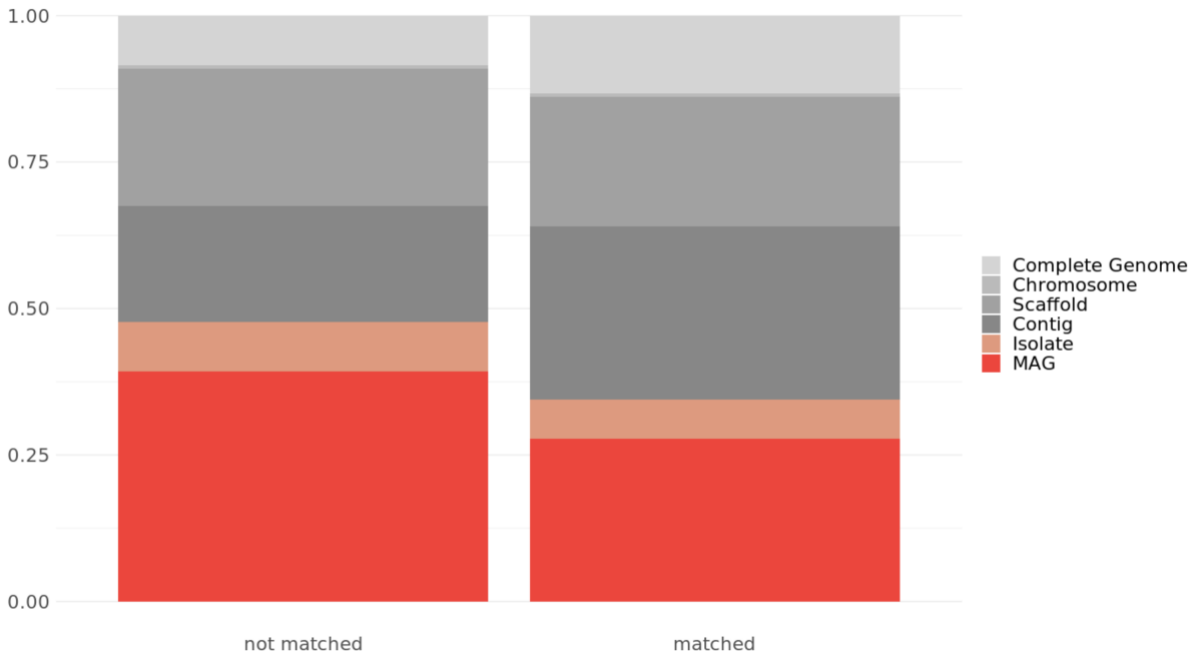

Figure 2. The figure shows which fraction of genomes are not matched and which are matched by GA-probes with the different genome categories as the mode genome category. There are two UHGG categories (MAG and Isolate) colored in red and four different RefSeq

categories in grey. The figure shows that the genomes not matched by the GA-probes have a substantially larger fraction of UHGG genomes than those matched.

Figure 2 shows that the genomes with a genome category from UHGG as a genome category (colored in red) make up a larger fraction of the not matched genomes than the matched ones. Nearly half of the not matched genomes have UHGG as the mode genome category, which applies to around one-third of the matched ones. This indicates that the assembly quality of UHGG might affect the results.

3.2.2 Matches between GA-probes and HumGut16S

The GA-map consists of 48 probes that are a complementary match to a stretch of the 16S sequence for some of the genomes in HumGut16S. Figure 3 shows how many probes match the 4253 genomes in HumGut16S. The leftmost bar shows that 1119 genomes have no probes matching. The bars to the right added up give 3134 genomes matched by at least one probe. Thus, almost three-fourths of the genomes are matched by at least one probe.

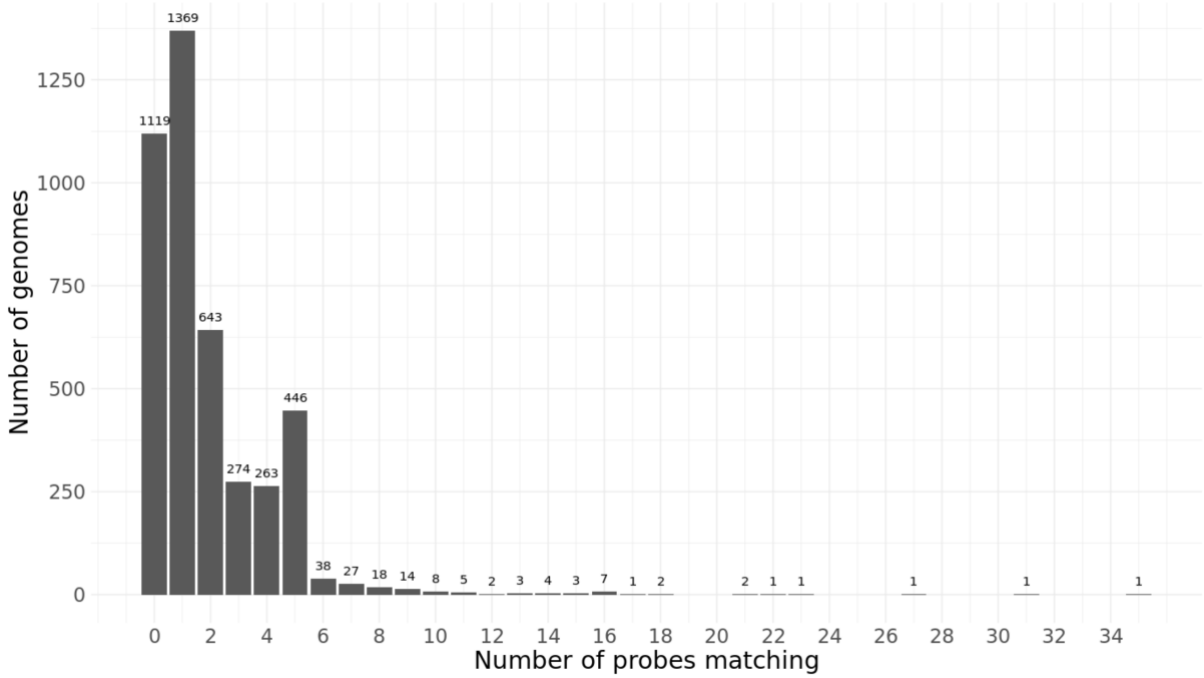

Figure 3. The figure shows how GA-probes matched the HumGut16S genomes, where the bars indicate how many genomes were matched by 0, 1, …, 35 different probes. Almost three-fourths of the genomes are matched by at least one probe.

Due to the differences in matching rate between different genome categories, two different analyses were performed: one including all genomes (figure 3) and one including RefSeq-genomes only (figure 4). Figure 4 shows the number of matches on the genomes in this subset of data, which contains 2936 genomes.

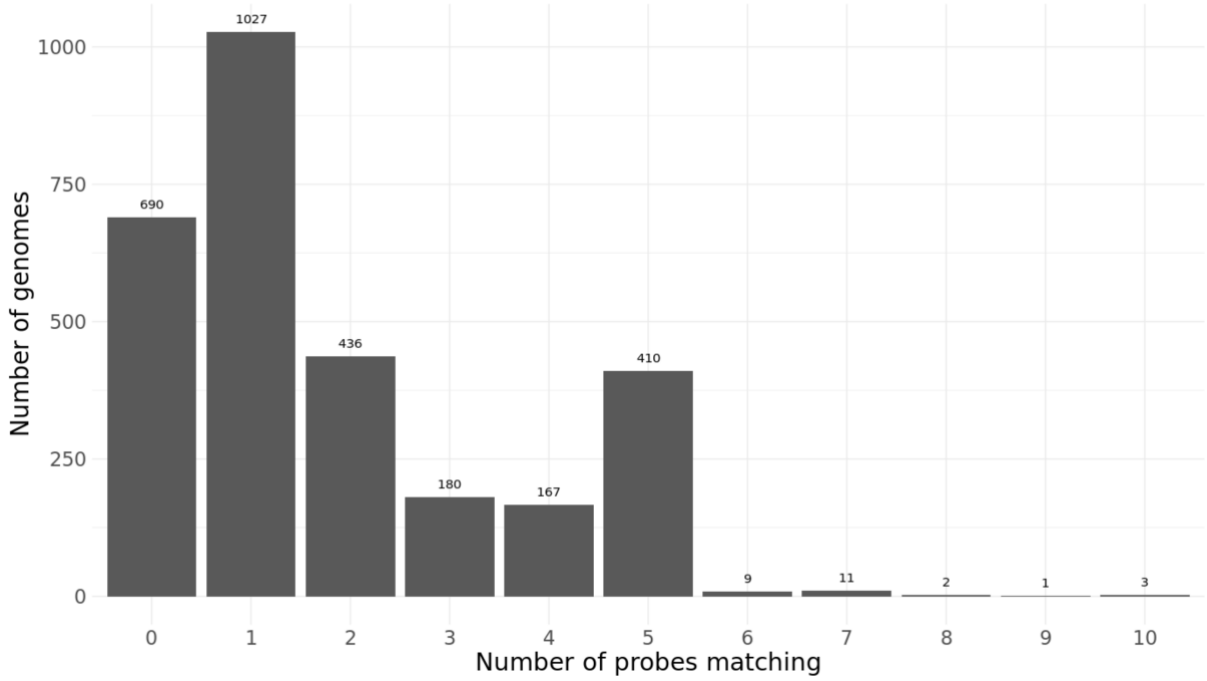

Figure 4. The figure shows how GA-probes matched the 2936 RefSeq genomes in HumGut16S, where the bars indicate how many genomes were matched by 0, 1, …, 10 different probes. Above three-fourths of the genomes are matched by at least one probe. The maximum number of probes matching one genome is 10.

Figure 4 shows that one probe matching is the most common, but some genomes have up to 10 matches. The bars with one or more matches summed up shows that 2246 genomes are matched by at least one probe, while 690 have no match.

Of 15 phyla present in HumGut16S, four are not matched by the probes. These are *Elusimicrobia*, *Synergistetes*, *Lentisphaerae*, and *Candidatus Saccharibacteria*. They are all among the six phyla with the fewest genomes in HumGut.

In the subset of HumGut16S with RefSeq-genomes only, there are 256 genera represented. Of these, 81 have no genome that is matched by the probes. Twenty-nine of these were found in under 1 percent of the healthy human guts that were screened to make HumGut, while two of them, *Ruthenibacterium* and *Agathobaculum*, were found in over 22 % of these guts.

3.2.3 GA-map probes

Matching between genome and probe was also analyzed on probe basis. Figure 5 shows how the probe matches on RefSeq-genomes are distributed across clusters and the taxonomic ranks of species and genera.

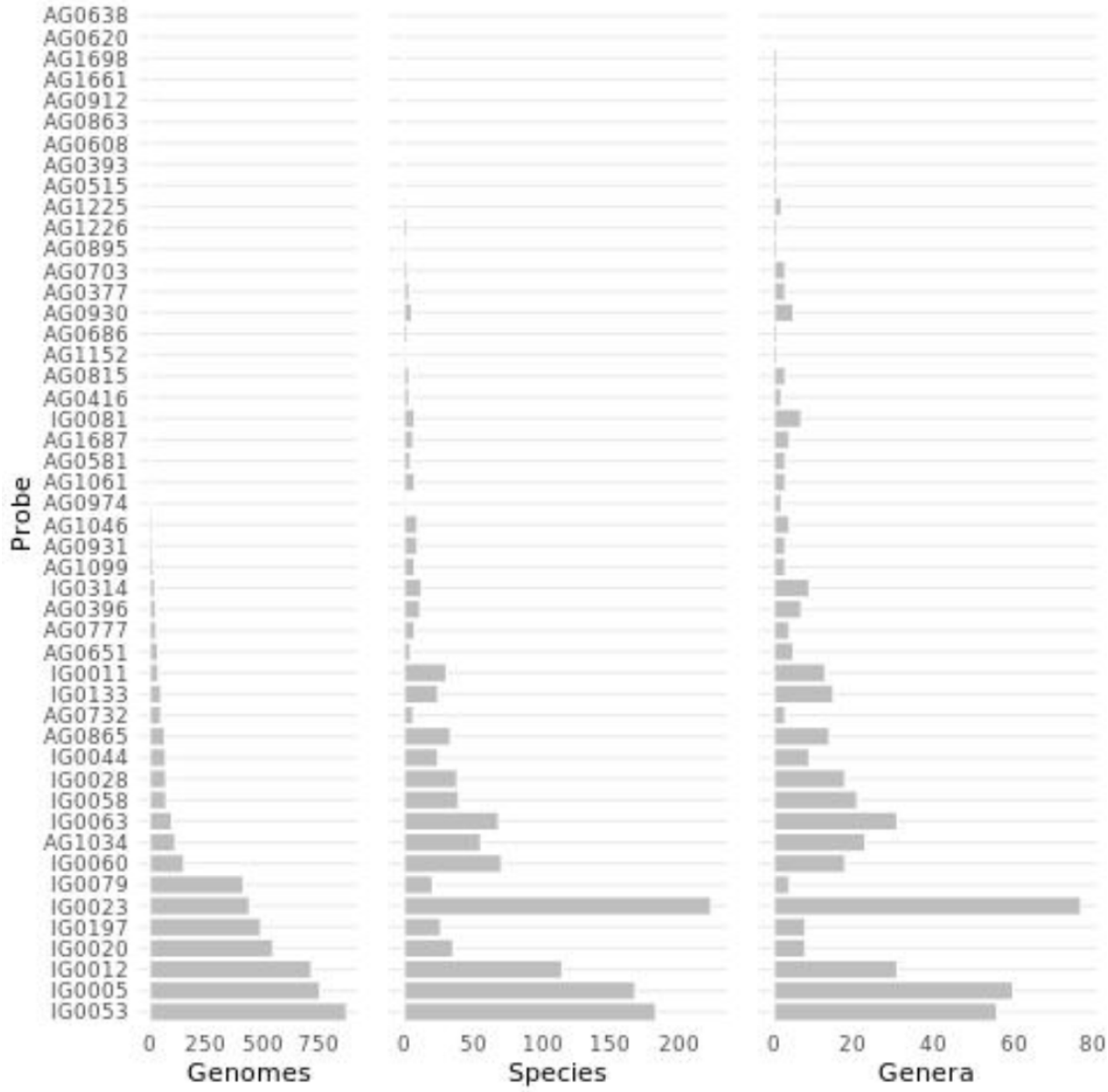

Figure 5. The figure shows how many genomes, species, and genera the different probes match using RefSeq genomes only. Two of the probes have no match against RefSeq-genomes.

The figure shows that some probes are narrow-matching and match only a few genomes. These are the upper probe names, and most match a narrow group of species and genera.

Some probes, like IG0023 and AG0930, match a wider group of genera than probes matching the same number of clusters. Two probes, AG0638 and AG0620, have no match against RefSeq genomes.

The company gave the GA-map probes anonymous names to hide their identity. These names do not contain any obvious information about the probes. The probes were therefore named after the genomes they match. This naming shows which narrowest classification level, over 75 % of the genomes the probe matches belong to, using RefSeq observations only. Table 3.1 shows the assigned names of the probes.

Table 3.1. The table shows the names of the probes given after the narrowest taxonomic group that over 75 % of the RefSeq matches belong to. Probes that did not get a unique name have added the number of clusters they match.

f_ Enterobacteriaceae , 40	IG0011	g_ Parabacteroides	AG1099	s_ Anaerobutyricum hallii	AG0608
f_ Enterobacteriaceae , 52	IG0133	g_ Pediococcus	AG1061	s_ Bacteroides fragilis	AG0377
f_ Enterobacteriaceae , 756	IG0005	g_ Streptococcus , 417	IG0079	s_ Bacteroides stercoris	AG0416
f_ Peptostreptococcaceae	IG0058	g_ Streptococcus , 52	AG0732	s_ Bacteroides uniformis	AG0863
f_ Veillonellaceae , 15	AG0931	Not targeting any	AG0638	s_ Clostridium sp.	AG0515
f_ Veillonellaceae , 72	IG0044	Not targeting any	AG0620	s_ Coprobacillus cateniformis	AG1661
g_ Akkermansia	AG0815	p_ Actinomycetota	IG0314	s_ Faecalibacterium prausnitzii	AG0651
g_ Alistipes	AG1226	p_ Bacillota , 100	IG0063	s_ Massilioclostridium coli	AG0912
g_ Bacteroides , 153	IG0060	p_ Bacillota , 115	AG1034	s_ Mycoplasma hominis	AG1698
g_ Bacteroides , 29	AG0396	p_ Bacillota , 12	AG0974	s_ Oscillibacter sp.	AG0393
g_ Bifidobacterium , 69	AG0865	p_ Bacillota , 444	IG0023	s_ Parabacteroides merdae	AG0686
g_ Bifidobacterium , 74	IG0028	p_ Bacillota , 717	IG0012	s_ Phascolarctobacterium faecium	AG1687
g_ Catenibacterium	AG0895	p_ Bacillota , 874	IG0053	s_ Ruminococcus bromii	AG1152
g_ Dialister	AG0930	p_ Proteobacteria	AG0777	s_ Streptococcus agalactiae	IG0081
g_ Dorea	AG0581	s_ [Ruminococcus] gnavus	AG0703	s_ Streptococcus pneumoniae , 494	IG0197
g_ Lactobacillus	AG1046	s_ Alistipes onderdonkii	AG1225	s_ Streptococcus pneumoniae , 548	IG0020

The table shows that 18 probes are named after species, 14 after genera, six after families, eight after phylum, and two observations do not match any RefSeq member. Probes named after species are generally more narrow matching than those named after phylum.

3.3 Functional profiling

The first step in functional profiling of the genomes in HumGut is to build functional profiles for each genome. In this project, this means:
1. Predicting genes with the prodigal tool and filtering them
2. Aligning the resulting proteins against Tax4FUN2 using diamond
3. Achieving the Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologs belonging to these proteins and filtering them

After the profiles have been built, the resulting functional profile for each genome is the vector of KEGG orthologs (KOs), indicating the presence or absence of the functional categories within the genome. The vectors used in this project are 0 or 1, where 1 means that at least one gene possesses this functional category.

In this section, results from HumGut16S will be presented first, followed by results from all HumGut-genomes. All genomes have been processed equally. When it comes to matching by probes, only genomes in HumGut16S can be considered matched.

3.3.1 Building functional profiles

Functional profiling is about finding the coding genes in each genome and then running a search with these against some database to assign the genes into functional categories. The functional profile for each genome is then a vector with 0 and 1, indicating which categories were found in that genome. When building functional profiles, one must decide upon different limits. While no limits are “correct”, these choices are important for the resulting functional profiles.

3.3.1.1 Gene prediction

The software Prodigal was used to predict coding genes in the genomes. This is needed to find functional categories within the genes. Gene prediction software tends to be over-sensitive and find genes that do not exist. For each predicted gene, the prodigal gives a score. The higher this score is, the more likely it is that the gene is real. In this case, each genome

contains many genes. Thus, being conservative and losing some genes is better than getting false positives.

To find a reasonable score limit for when to believe that a gene is real, three random DNA-"genomes" similar to the genomes were made. These do not contain any actual genes. Thus, any genes predicted by prodigal are false positives for these "genomes." The histograms in Figure 6 show the prodigal scores of the three randomized DNA "genomes" and three cluster representatives from HumGut. The plots show that the score for random genomes is generally under 10, while the score of the cluster representatives is generally below 1000.

A score limit of 30 was decided by comparing the random DNA and cluster representatives. This limit is shown with a red line in Figure 6. All predicted genes below this limit were discarded.

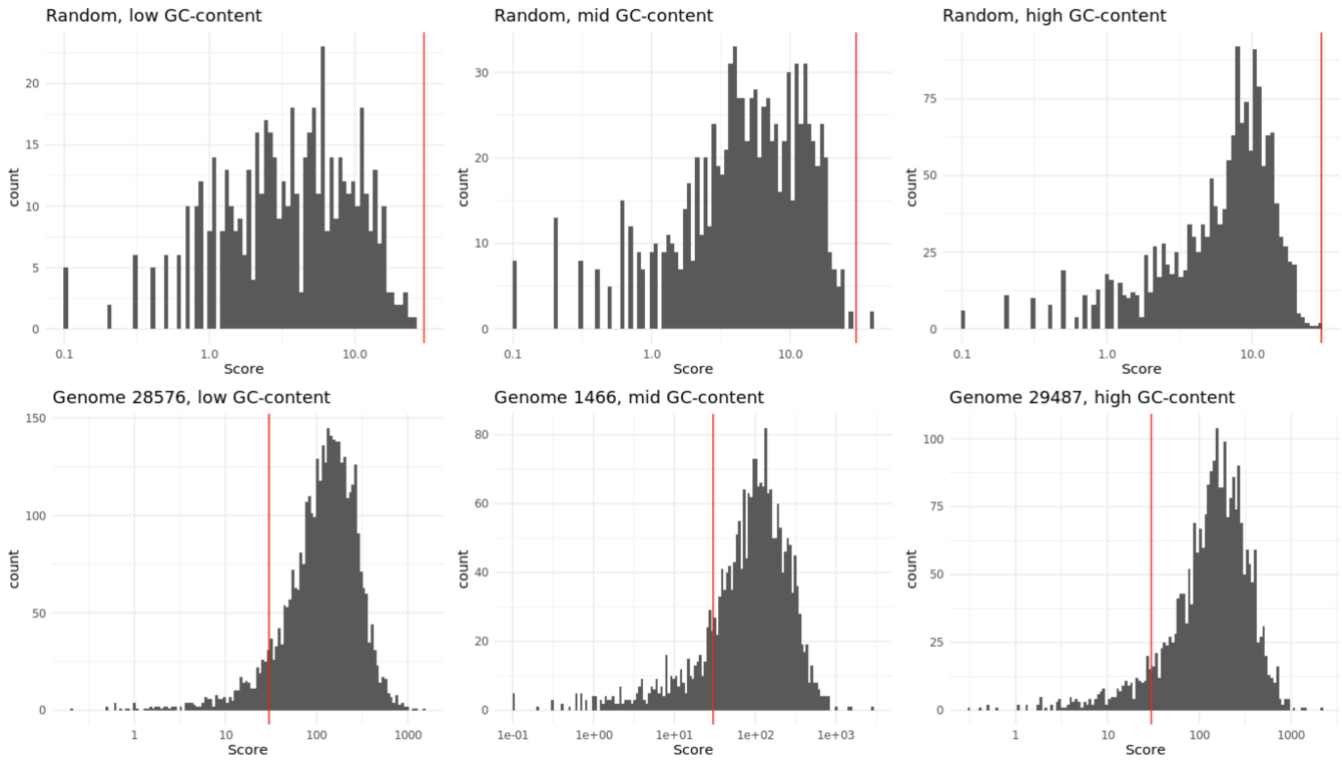

Figure 6. The figure shows scores for genes predicted by prodigal for three random DNA "genomes" (top) and three genomes with different GC content. The red lines show the decided score limit of 30. All predicted genes below this score were discarded.

3.3.1.2 KEGG Orthologs

Functional profiling was performed for the clusters in HumGut16S and the HumGut-dataset. The software Diamond was used. Diamond searches after alignments in the Tax4FUN2-database. The database contains 6 323 861 proteins assigned to functional KOs. There are in total 21 620 KOs. 8236 of these are present in the coding genes predicted by HumGut16S. The histogram in Figure 7 shows the distribution of how many cluster representatives the KOs were found in.

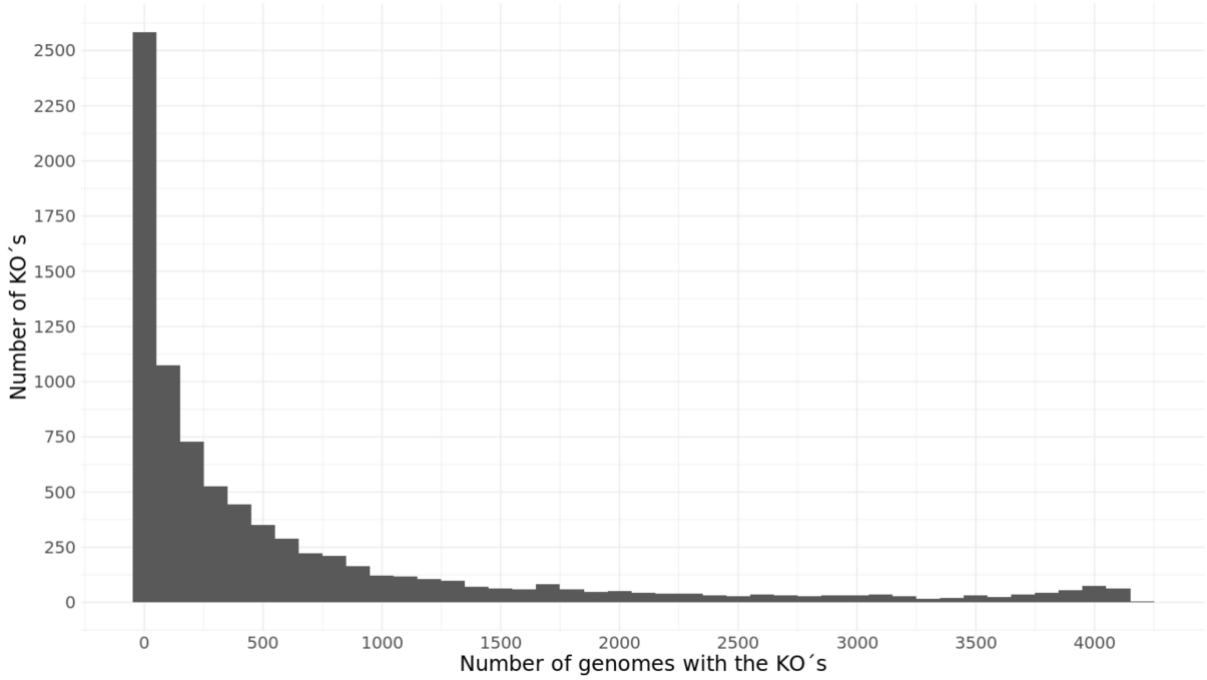

Figure 7. The figure shows how KEGG orthologs (KO) distribute across HumGut16S. As shown to the left, most KOs are found in very few genomes. The rightmost bars show that some KOs are found in almost all genomes. No KO is found in all genomes.

In this project, Diamond results with an e-value above 0,001 were discarded. This means that for each 1000 search, it is expected to get one false positive. These KOs are most likely to be occurring infrequently. 872 functional categories were found in three or fewer genomes and were removed. The 7364 remaining functional categories were found in between 4 and 4178 of the 4253 genomes.

In the full HumGut-dataset, including HumGut16S, 8481 functional categories out of 21620 were found. The most frequently occurring functional category was found in 29129 of 30536 genomes. The histogram in Figure 8 shows how many cluster representatives the KOs were found in.

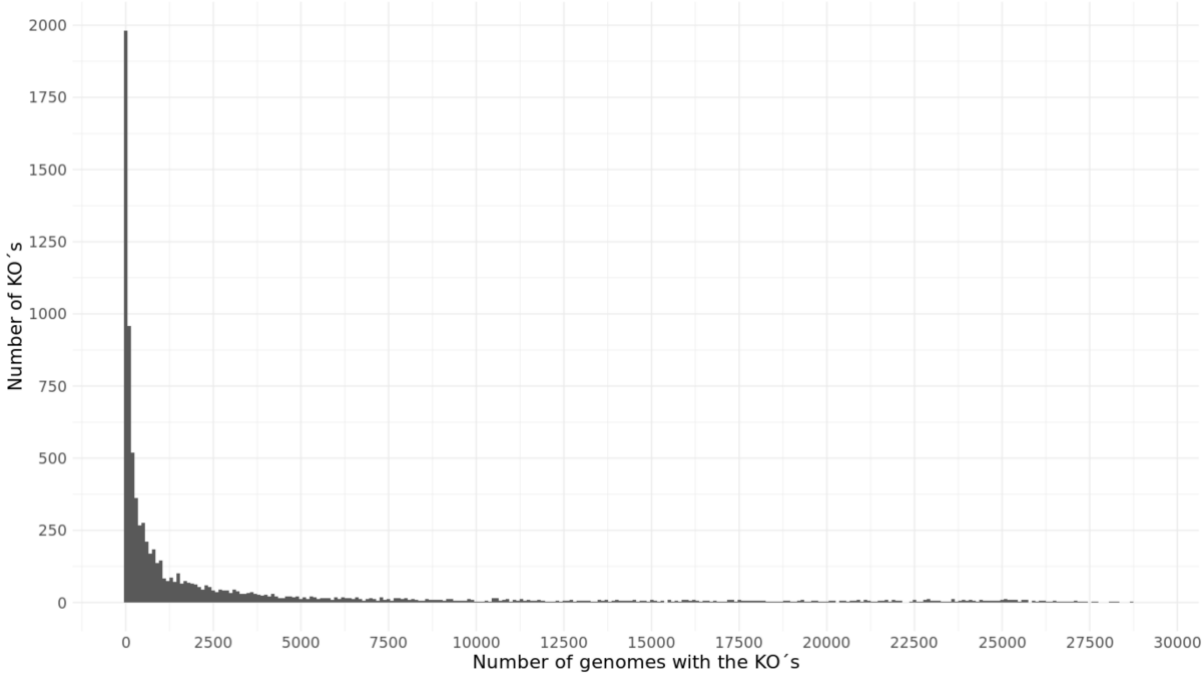

Figure 8. The figure shows how KEGG orthologs distribute across all the HumGut-genomes. As shown to the left, most KOs are found in very few genomes. The rightmost bars show that few KOs are found in most genes. No KO is found in all genomes.

Like explained above, it is expected to observe some false positives. An E-value threshold of 0,001 was used for both analyses. In HumGut, 316 KOs were found in three or fewer genomes and were discarded. Thus, 7925 functional categories were kept.

3.3.2 Analysis of the functional profiles

Functional profiling can be used to find patterns in the function between different categories of genomes. One way to analyze functional profiles is using Principal Component Analysis (PCA). As explained in the method section, PCA is a tool that reduces multidimensional data into components, whereas as few components as possible explain as much of the variance in the data as possible (Daffertshofer et al., 2004). PCA plots show the two variables that explain the most variance, named PC1 and PC2. In this project, functional profiling is used to

compare the different assembling sources and to investigate how the GA-map spans over the functional space of the human gut.

3.3.2.1 Using functional profiles to separate genomes by phylum

HumGut contains 16 phyla (HumGut16S contains 15). Genomes within the same phylum are expected to share a lot, even the majority, of the functional categories. Visualizing how a PCA separates the phyla can reveal whether two principal components are enough and show which genomes that are separated in later PCA plots.

Figure 9 shows how well PC1 and PC2, which are the two components explaining most of the variance, can separate the different phyla in HumGut16S. The two principal components make up nearly one-fourth of the found variance in the profile. The phylum list on the right side of the plot is sorted after the number of genomes within the phyla in HumGut16S, from most genomes to fewest.

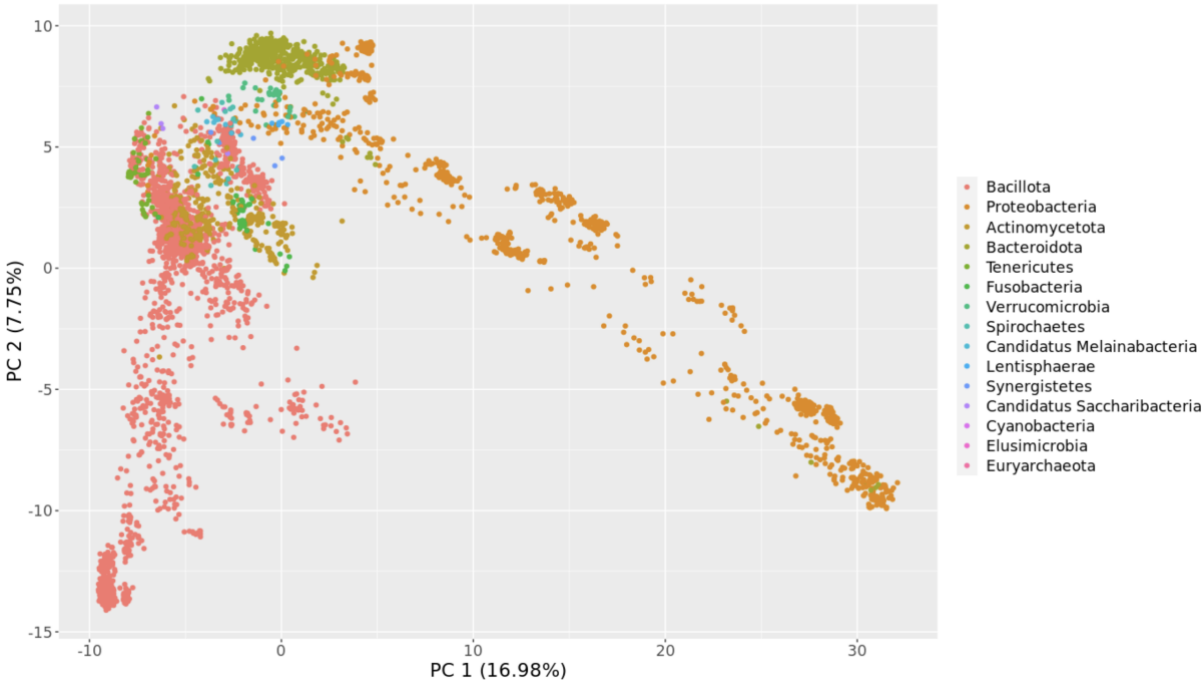

Figure 9. The plot visualizes a principal component analysis for HumGut16S, where the two first PC´s explain 24.73 % of the variance. Different colors are used to show which phylum the cluster representative belongs to. The phylum list on the right side of the plot is sorted after the phyla´s frequency.

By using two components only, some separation can be seen between the 15 phyla. For instance, *Proteobacteria* generally have a higher value on PC1, while *Bacillota* is generally low on PC1 and has a wide range of PC2. *Bacteriodota* is mainly gathered at high PC2 and medium PC1.

The same analysis was performed for all the genomes in HumGut, shown in Figure 10. As in Figure 9, the phylum list is sorted after the number of genomes in the figure.

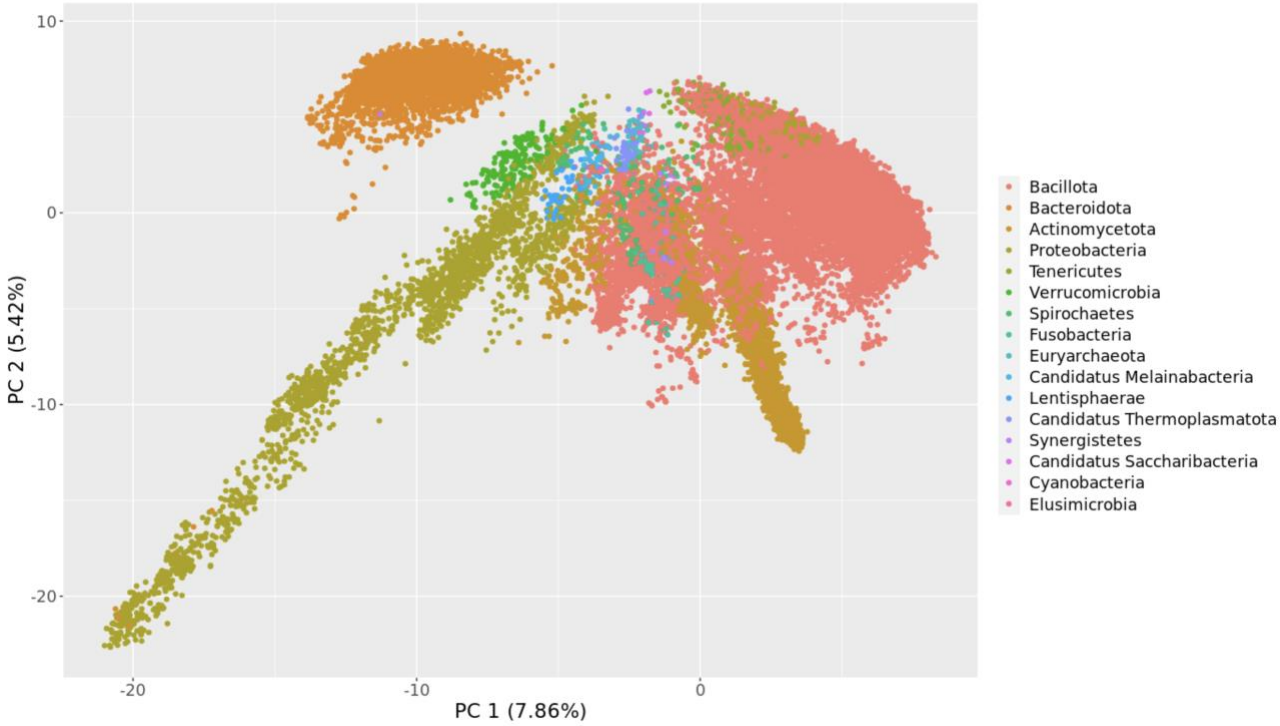

Figure 10. The plot visualizes a principal analysis for HumGut, where the two first PC´s explain 13,28 % of the variance. The phylum list on the left is sorted after the phyla´s frequency. Different colors are used to show which phylum the genome belongs to.

The two different principal components explain 13.28 % of the variance found. These two components show a clear separation between the phyla. As seen in Figure 10, showing the phyla separation from functional categories in HumGut16S, most *Proteobacteria* genomes are clearly separated from the others. *Bacteriodota* seems to be clearly different from the others.

3.3.2.2 How HumGut16S is distributed in HumGut

As mentioned earlier, there are in total 30536 clusters, and 4253 of these are in HumGut16S. Figure 11 is a PCA plot, explaining a total of 13.28 % of the variance in HumGut. In the figure, genomes in HumGut16S are colored red, and the rest of HumGut is blue.

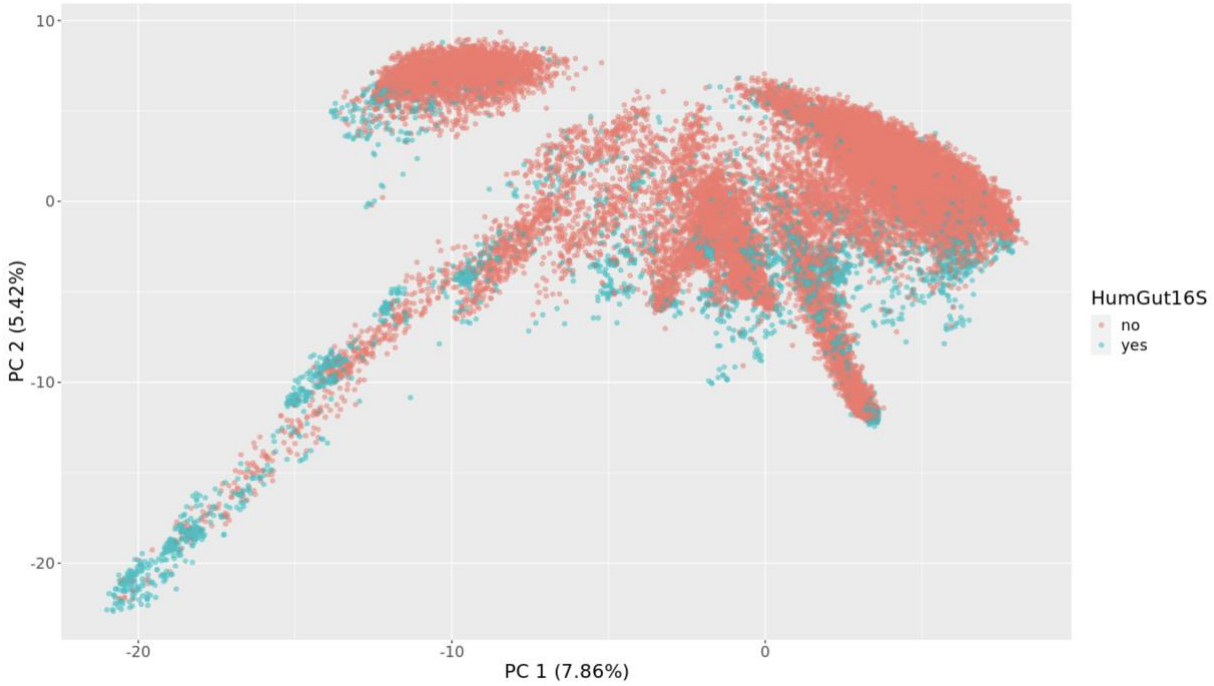

Figure 11. The figure shows a PCA plot that explains 13.28% of the variance found in the functional profiles of HumGut. The 4253 in HumGut16S are colored blue, while the 26283 other genomes are red. The plot shows no clear separation of HumGut16S and the rest of HumGut, which indicates that no functional region is not partially in HumGut16S.

Figure 11 shows no distinct differentiation between genomes that are present in HumGut16S and those that are not. This indicates that there are no regions in functional space that are not partially covered by HumGut16S. Thus, it seems like the HumGut16S can be used as representatives for HumGut.

3.3.2.3 Functional profiling and match of a GA-probe

Figure 12 shows how well the same two principal components can separate between genomes matched by and not matched by the probe IG0023. IG0023 was chosen as an example because it is the most wide-matching probe in terms of species and genera, shown in figure 5. For more narrow-matching probes, separation in principal components is more expected regardless of the probe.

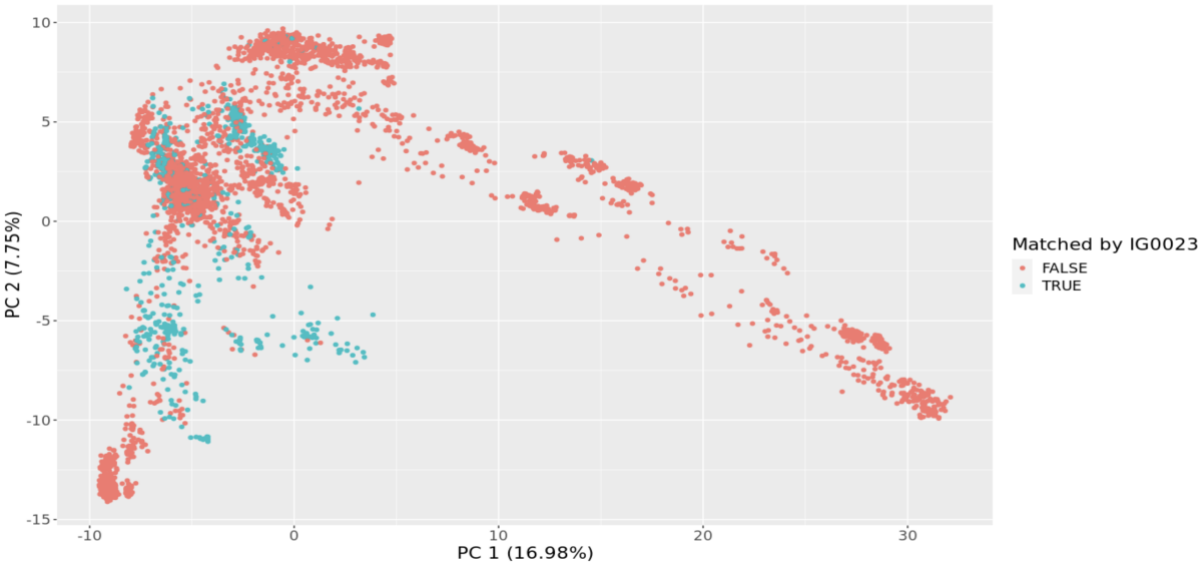

Figure 12. The figure shows a principal component analysis, where the two first principal components explain 24.73 % of the variance. IG0023 matches 905 genomes (colored blue in the figure). There are 3348 HumGut16S genomes not matched by this probe (colored in red).

Figure 12 shows that most of the IG0023-matched genomes are grouped, but there is no perfect separation. This indicates that most of the genomes matched by IG0023 are functional similar, and that also some other genomes also share this similarity. IG0023 matches 905 genomes. This is a greater number than shown in Table 3.1 and Figure 5, which show RefSeq-genomes only.

3.3.2.4 Variation in functional profiling between different genome categories

The HumGut16S contains six genome categories: four RefSeq (complete genome, chromosome, scaffold and contig) and two UHGG (isolate and MAG). 1310 genomes are MAG, 1148 contig, 953 scaffold, 511 complete genome, 304 isolate, and 27 chromosome. Only RefSeq-genomes are curated, and if this affects this analysis, the two different sources are expected to be grouped separately in PCA plots. Figure 13 shows two PCA plots showing how the different genome types are spread. The left bar shows the four RefSeq categories, and the right panel UHGG. The figure is split to make the categories with few genomes visible.

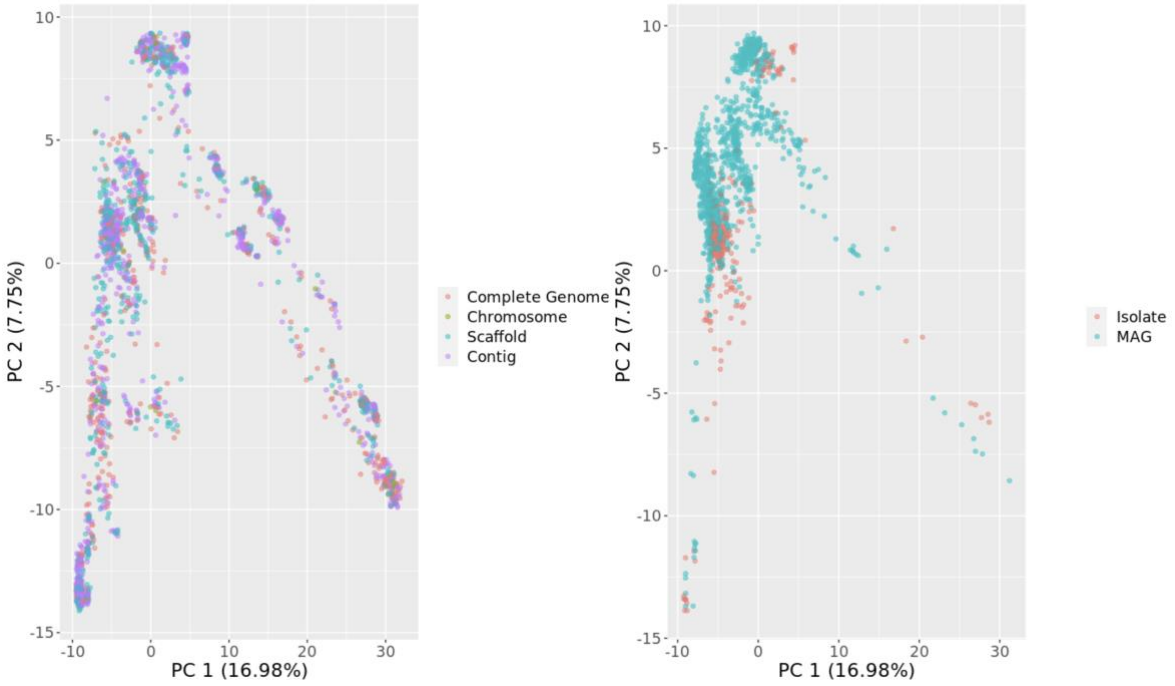

Figure 13. The figure shows how the genome categories in HumGut16S are distributed across the two principal components. The left panel contains RefSeq categories, which are complete genome (511 genomes), chromosome (27 genomes), scaffold (953 genomes) and contig (1148 genomes) and the right panel UHGG, which contains 304 isolate genomes and 1310 MAG . The figure is split to make all categories visible.

Figure 13 shows some vague grouping for MAGs at PC2 between 0 and 10 and PC1 between around -8 and 0. PC1 around 0 and PC2 around -6 seem to be RefSeq genomes only. Still, it is not a good separation between the different genome categories.

Assembling of 16S-sequence in MAGs is difficult, so HumGut contains a larger fraction of MAGs than the 16S-dataset. Out of 30536 genomes, 27363 are MAGs, 1247 are contigs, 1041 are scaffolds, 512 are complete genomes, 343 are isolates, and 30 are chromosomes. Figure 13 shows two PCA plots that visualize how the different genome categories are spread in HumGut. The left panel shows RefSeq-categories, and the right UHGG. The figure is split into two panels to make all the categories visible.

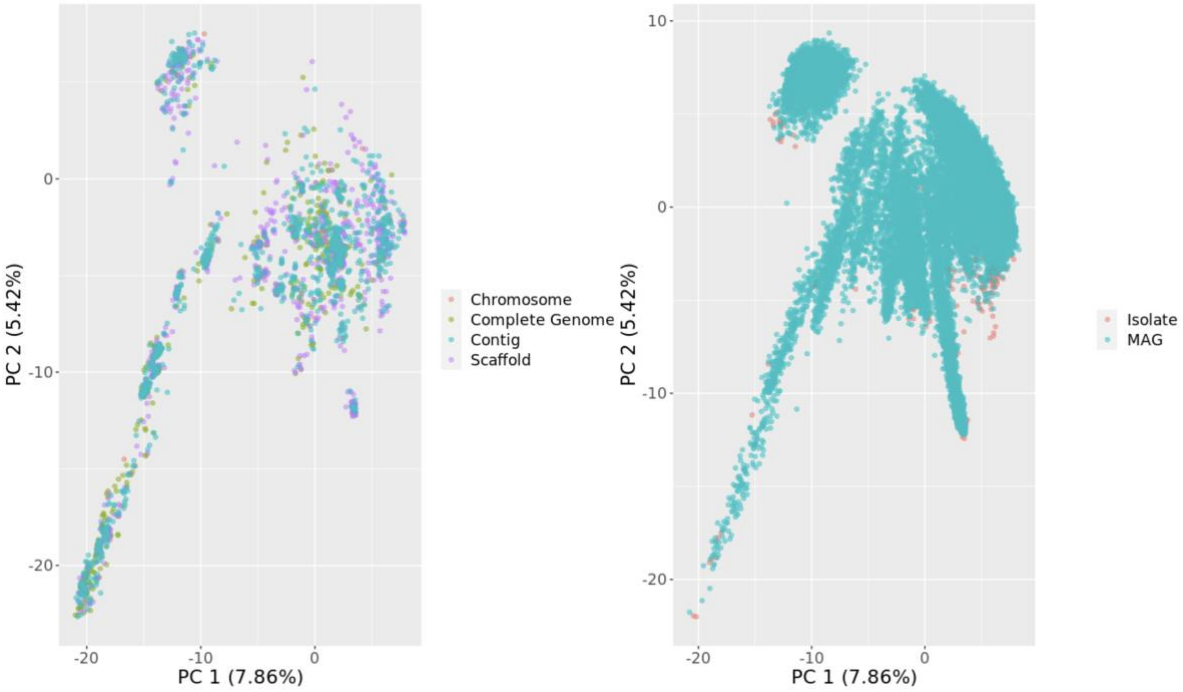

Figure 14. The figure shows how the genome categories are distributed across the two principal components. The left plot contains RefSeq categories (Chromosome, 30 genomes, complete genome, 512 genomes, scaffold, 1041 genomes and contig, 1247 genomes) and the right plot UHGG (isolate, 343 genomes and MAG, 27363 genomes). The figure is split to make all categories visible.

This figure shows that some regions have higher RefSeq frequency than others. Most genomes with PC2 lower than -15 are RefSeq. No large region seems to be entirely missed by any of the sources. Even if some grouping can be seen, PCA does not separate the different genome types well.

The small trends of spreading of genome categories in Figures 13 and 14 might be because the functional profiles differ between the different phyla. Figure 15 is a PCA plot showing the source (RefSeq or UHGG) for *Bacillota*-genomes only. *Bacillota* makes up over half of the genomes in HumGut, with 17488 genomes. As stated earlier, the source is the mode source for genomes in HumGut16S and the cluster representative´s source for other HumGut genomes.

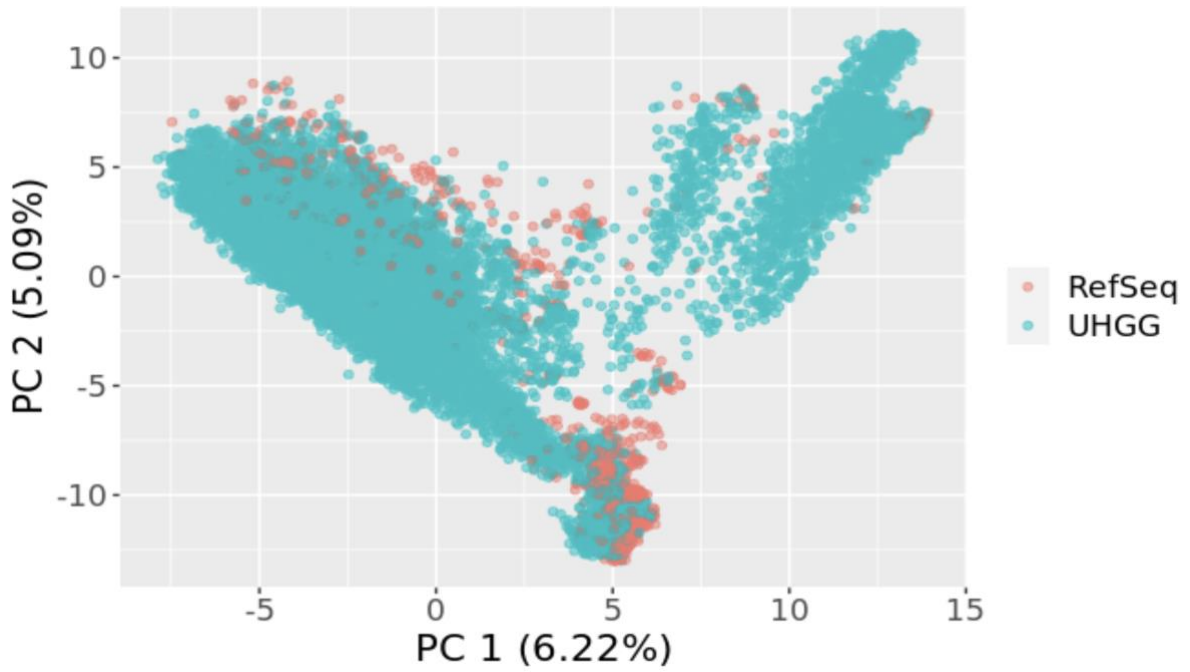

Figure 15. The figure shows a PCA plot for the phylum Bacillota, where the two PCs explain 11.31 % of the variance. The 1350 genomes from RefSeq as the source are colored red, while the 18217 UHGG genomes are blue. A slight pattern between RefSeq and UHGG can be seen for some of the genomes. For instance, the RefSeq observations at PC1 -5 tend to have higher PC2 than UHGG-genomes.

The PCA plot shows a slight trend to separation between RefSeq and UHGG. For instance, the RefSeq observations at PC1 -5 tend to have higher PC2 than UHGG-genomes. This pattern is vague, not perfect, and only seen for a few genomes. Thus, the PCA plot does not provide a good separation between the sources.

Using PLS is a different way to analyze functional profiles. PLS aims to maximize covariance between the response and other variables (Wold et al., 2001), and any differences between the sources should thus be visible in a PLS plot.

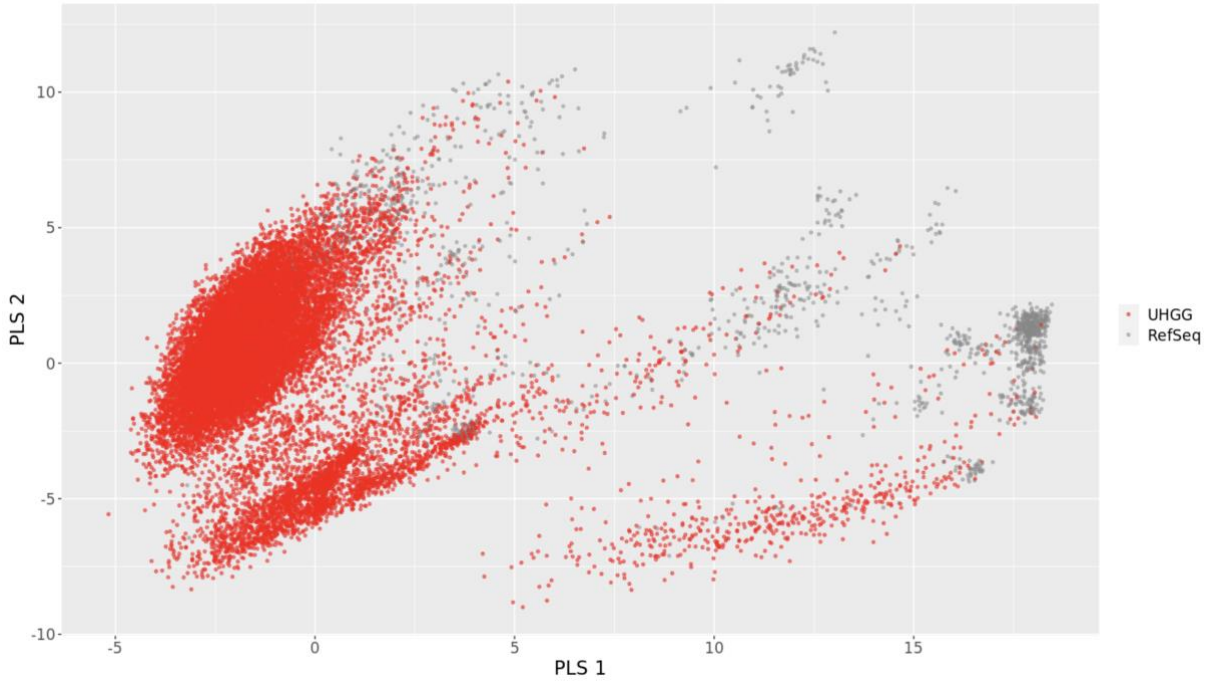

Figure 16. The figure shows a PLS plot for Bacillota-genomes in HumGut. The 18217 Bacillota-genomes with UHGG as the source are colored in red, while the 1350 genomes from RefSeq are grey. The plot shows some trends. Most genomes to the far right are RefSeq. The genomes with the lowest value on PLS component 2 tend to be UHGG, while genomes with PLS component 2 above 10 are mostly RefSeq. In total, it is still not a good separation between the two sources. Using the same data as test and training data gives an accuracy of 0.96.

To narrow the functional spread due to differences in the genomes further down, the genera *Streptococcus* was extracted and studied further. A PLS for this genus only is shown in figure 17.

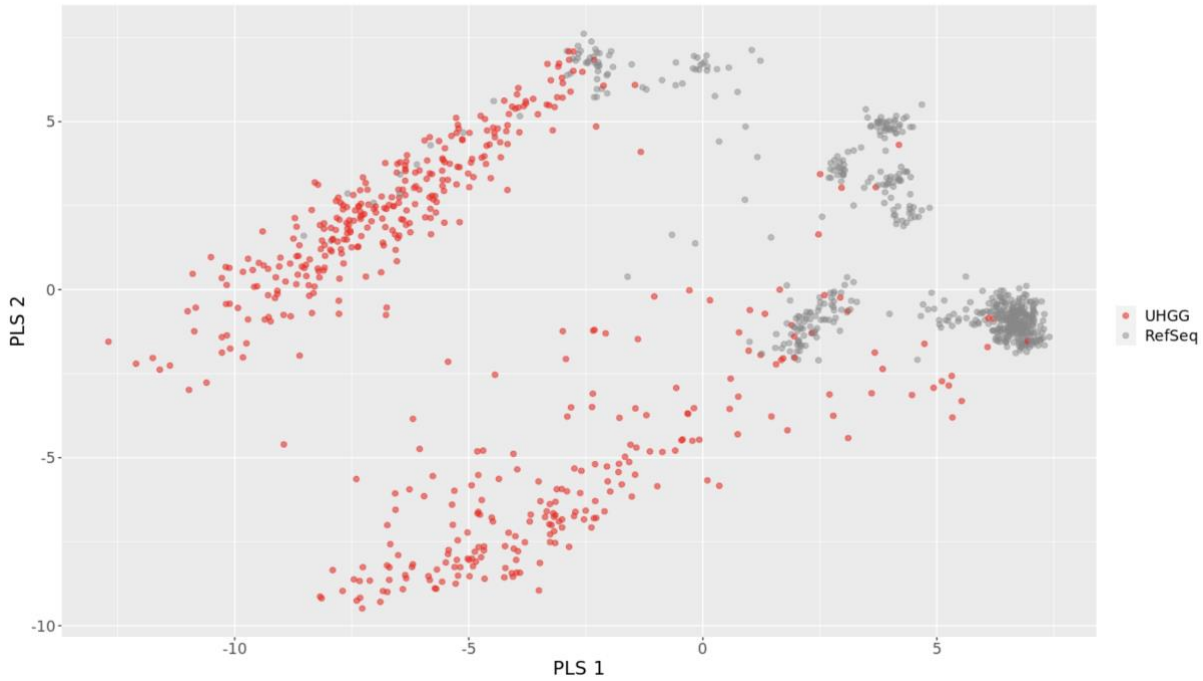

Figure 17. The figure shows a PLS plot for the Streptococcus genera. There are 505 from UHGG are colored in red, while 583 RefSeq-genomes are grey. The plot shows a good separation between genomes from the two different sources. Using the same data as training and test data gives an accuracy of 0.93.

Figure 17 shows a strong relation between PLS and source for one genus only. This might indicate that it is possible to find KOs, or combinations of KOs, that separate the two sources.

This was investigated further using correlation. Correlation between the functional categories and source was calculated for the 2927 KOs that had variation within *Streptococcus*. Two KOs were found in all *Streptococcus*-genomes, while 4997 were not found in any. The two highest correlations in absolute value were 0.75 and 0.74. These correlations belonged to functional categories found in the majority of RefSeq genomes and few UHGG genomes.

The two KOs with the highest correlations are K12294 and K12295. These are both associated with the spliceosome and are involved in processing genetic information at the transcription level. Both KOs are two-component systems and belong to the LytTR family. K12294 is a sensor histidine kinase, while K12295 is a response regulator.

There are nine KOs with an absolute correlation above 0.65. All of these are functional profiles that are more frequent in RefSeq than in UHGG-genomes. These are associated with metabolism, environmental information processing, genetic information processing, organismal systems and human diseases.

3.3.2.5 Clustering with K-means

3.3.2.5.1 Clustering HumGut16S with K-means

While PCA can give valuable insight into the data, it only explains a fraction of the variation in the dataset because it compresses everything into two dimensions. Clustering is to divide the data into groups with similar data based on all dimensions. Thus, each cluster is a "subset" of the functional space of HumGut, which may reveal variation not seen with PCA. Here, clustering is used to investigate further whether the GA-map covers all functional areas of HumGut and the differences between the genome assembly sources.

Using K-means clustering, the 4253 genomes in HumGut16S were assigned to 43 clusters, where the number 43 was chosen to have, on average, close to 100 genomes in each cluster. The clusters contained between 11 and 261 genomes. Figure 18 shows the fraction of genomes within each k-mean-cluster that is marched by the probes.

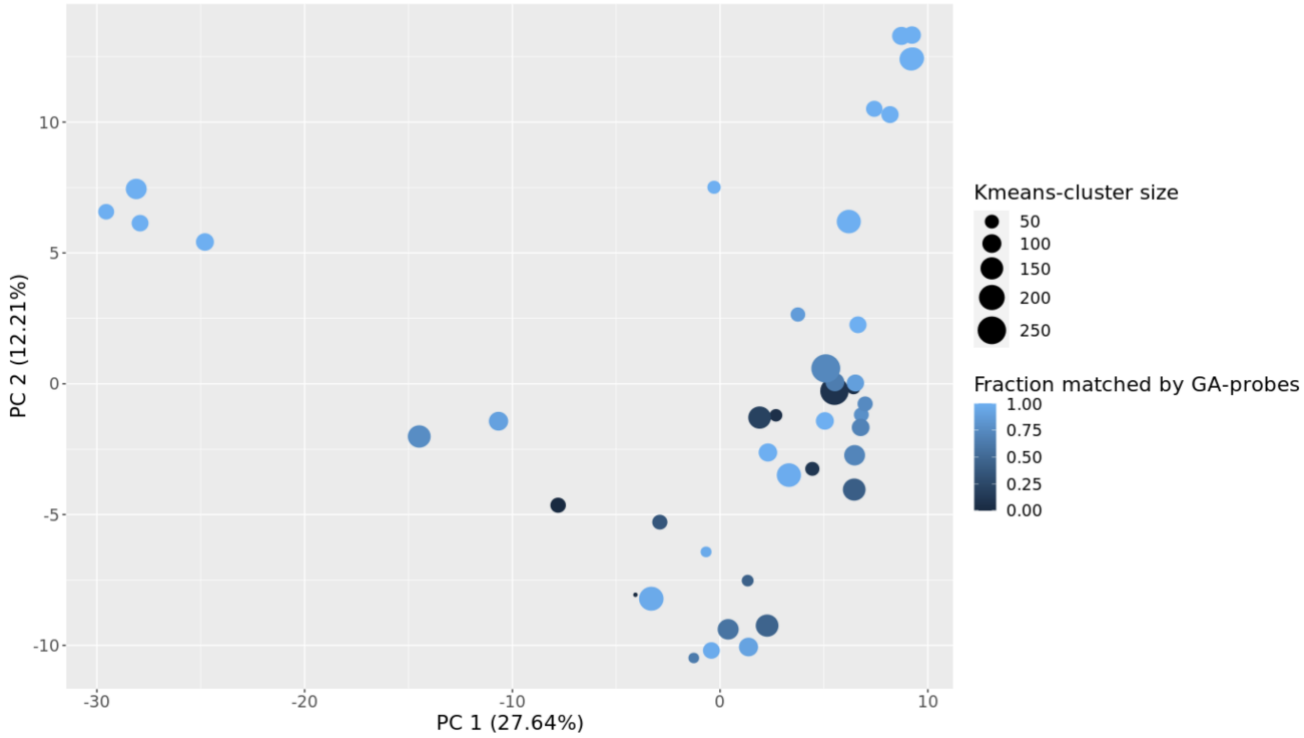

Figure 18. The PCA plot shows the 43 K-means clusters as dots. The size of the dots illustrates the size of the K-means clusters in terms of genomes belonging to the cluster. The color illustrates the fraction of the K-means cluster that the GA-probes match.

Some of the dots in Figure 18 have a dark color. Of these, two are not matched by any GA-probe. The 15 dots with the lightest color illustrated the 15 K-means clusters where GA-probes match all members. The plot shows that most of the colors with a light color have a high PC2.

The K-means clusters that the GA-probe does not match contain 63 functional categories that are not present in the fully matched clusters. Among these, seven categories are not matched at all. Two of these are only found in the non-matched clusters. There are, in total, 60 functional categories not found in matched genomes. These are, on average, found in 7.73 genomes, while the average is 696 for all functional categories in HumGut16S.

The two clusters not matched by the GA-probes are the two darkest dots in the plot. One of them is located at PC1 around -8 and PC2 -6. This dot represents 62 genomes, all belonging to the phylum Proteobacterium. The other not matched cluster can be seen as a tiny dot at PC1 -4 and PC2 -6, next to a bigger, lighter dot. These two clusters are both consisting of *Proteobacteria*-genomes. The second non-matched cluster is the smallest one and consists

of eleven genomes. The larger, lighter one next to it represents 179 genomes, where the probe matches 97 % of the members.

The smallest non-matched cluster has one functional profile found in all the members of this K-means cluster but not in any other. This functional profile is K21486, the KO identifier for the protein ankyrin repeat family A member 2, mainly known as ANKRA2 that enables enzyme binding activity in human cells (https://www.ncbi.nlm.nih.gov/gene/57763). There are also 15 other functional profiles that are found in over 90 % of the member of this non-matched cluster that is rarely found in other clusters. Among these, the functional profile occurring most frequently in other clusters is found in 14 % of its members.

The larger unmatched cluster has no functional profiles that are never found in matched clusters. Five functional profiles found in above 80 % of this cluster's members are found in a maximum of 20 % of the members of other clusters. One of these is only found in one genome outside of this cluster. This functional profile is related to biotin metabolism.

Two clusters with slightly lighter colors than the not-matched ones can be seen at PC1 around 2 and PC 2 around -1.2. The largest of them represents 151 genomes, belonging to the phyla *Actinomycetota*. The smallest one represents 40 genomes, all *Fusobacteria*.

At the left and upper right corners, there are groups of clusters that are lighter colored. The left group, showing four dots at PC1 below -20, are all fully matched by the probes. This group represents, in total, 353 genomes. Most of these belong to the phylum *Proteobacteria*, but there are also eight *Bacteriodota*-genomes. There are six clusters located at PC2 above 10. This group represents 146 genomes, all belonging to the phylum *Bacillota,* and all matched by the GA-probes.

3.3.2.5.2 Clustering HumGut with K-means

HumGut was assigned to 305 clusters using K-means. As above, the number of clusters were chosen to have, on average, 100 members in each. Figure 19 shows how these clusters are spread in size and fraction matched by the probes.

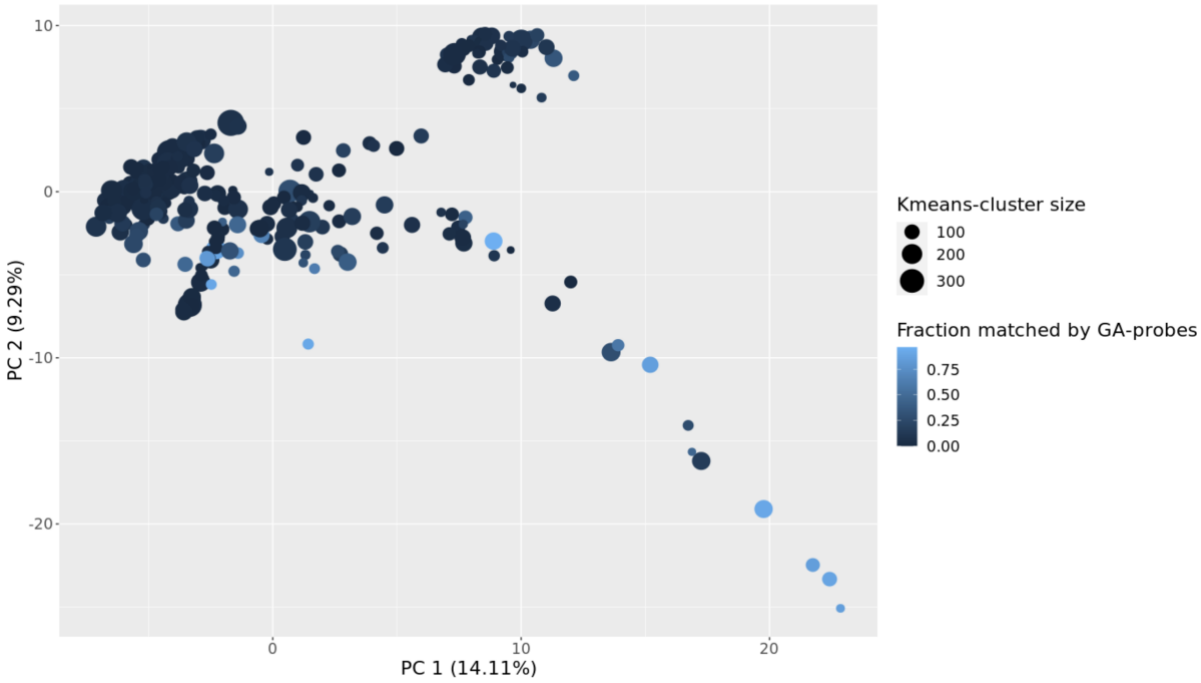

Figure 19. The PCA plot shows the 305 K-means clusters as dots. The size of the dots illustrates the size of the K means-clusters in terms of genomes belonging to the cluster. The color illustrates the fraction of the K-means cluster that is matched by the GA-probes. There are only HumGut16S genomes that can be analyzed for matches against the probes, and there are relatively few HumGut16S genomes in HumGut. Thus, most dark clusters have no or few HumGut16S members.

Figure 19 shows how well the probes match these 305 K-means clusters. The K-means clusters contain between 11 and 373 genomes. 66 clusters are not matched by any of the probes. Half of these, 33 clusters, have no genomes in HumGut16S and thus no possibility to be matched by the probes. 83 clusters have less than one percent of its genomes matched by the GA-probes.

Some groups in Figure 19 stick out. For instance, all the genomes below PC1 -5 seem dark and thus have a small fraction that is matched by the probes. This region of the PCA plot shows 56 K-means clusters, representing 6086 genomes. All of these genomes belong to the phylum *Bacillota*. The cluster with the highest matched rate from this group, has 0.29 of its members matched by GA-probes. This cluster is located at PC1 -5.2 and PC2 -4.1. The cluster located furthest to the left of the figure has a matched rate of 0.078.

Another group that is separated from the others in Figure 19 is located at PC2 above 5 and PC1 between 5 and 15. This group contains 44 clusters representing 4010 genomes. All of these belong to the *Bacteroidota* phylum except for one that is *Synergistetes*. The cluster has a mean fraction matched by the probes at 0.08, but the most matched cluster has 0.037. This genome is located at PC1 12.1 and PC2 6.98. Figure 19 shows that within this group, especially genomes with PC1 below approximately 8 seem to be dark. In this subgroup, there are nine clusters (858 genomes), and the highest matched rate is 8.3%.

The lower left corner in Figure 19 seems to have more clusters that are matched at a higher rate than the other areas in the plot. Four clusters have PC1 above 18. These have a matched rate of 0.85-0.92. The four clusters represent 358 genomes. Nine of these are *Bacteroidota*, while the rest are *Proteobacteria*.

4. Discussion

This master thesis aimed to investigate and make functional profiles of the genomes in the HumGut database, a database for microorganisms found in the human gut. The thesis had three aims: to gain a taxonomic and functional overview of HumGut, assess the coverage of the GA-map dysbiosis test from the company Genetic Analysis (GA), and evaluate the quality of the HumGut-genomes. First, this was done by investigating how the GA-probes match the subset of HumGut with available 16S sequence. This subset of HumGut will be referred to as HumGut16S. Secondly, functional profiles were made for all genomes found in HumGut, and different groups in the form of phyla, matches with probes or not, categories and sources were compared. The following discussion will first focus on classifying and then on functional profiles.

4.1 HumGut

There are 16 phyla in HumGut and 15 in HumGut16S. Figure 1 shows the number of genomes belonging to each phylum, both for HumGut and for HumGut16S, the latter colored green. The bar for *Candidatus Thermaplasmatota* has no green part and is not a part of HumGut16S. This phylum has 36 genomes in HumGut and is the fifth smallest phylum, as shown in the figure.

All the other phyla are present in HumGut16S. However, there are variations in what degree each phylum is presented. For some, one-third of the HumGut-genomes are in HumGut16S, but other phyla, like *Euryarchaeota*, are barely present. This variation makes it challenging to determine whether HumGut16S is a good representative subset for HumGut. This will be explored later.

HumGut consists of 30536 genomes, and 4253 of these are in HumGut16S. 16S is the most used marker for studying microbiomes (Pollock et al., 2018), but there are quite few HumGut-genomes with available 16S sequences. The lack of this sequence mostly comes from difficulties in assembling 16S from short reads. However, this might soon improve.

As explained in the introduction, 16S sequences are highly conservated. 16S-sequences are challenging to assemble from short-read MAGs for several reasons. First, because the 16S

contains repetitive regions (Perisin et al., 2016), it can be difficult to determine the correct order of the reads. Second, some genomes contain several 16S-sequences with slight variations (Větrovský & Baldrian, 2013). Because 16S-sequences are conservated, this sequence is often similar between different organisms (Yuan et al., 2015), which makes it challenging to determine which reads belong together. These challenges lead to few HumGut-genomes with available 16S-sequences and might result in poorer quality from HumGut-MAGs 16S sequences. However, if long-read MAGs are used instead, one read might cover the whole 16S sequence, and the problem of assembling this sequence is removed. A study from 2016 showed that Nanopore long reads could identify more species than short-read sequencing (Shin et al., 2016). This indicates that Nanopore long reads had higher quality and thus might be matched more similarly to RefSeq genomes than short-read MAGs. A study from 2022 showed that long-read MAGs had twice as many high-quality MAGs as short-read MAGs (Gehrig et al., 2022). Even though all MAGs in HumGut are considered to have high quality, this might mean that long-read MAGs could gain even higher quality.

4.2 How the GA-map overlaps HumGut

4.2.1 Genome categories

HumGut16S contains genomes assembled from two different sources, the National Center for Biotechnology Information (NCBI) Reference Sequences (RefSeq) and The Unified Human Gastrointestinal Genome (UHGG). These two sources contain a total of six different genome categories: MAG and Isolate (both UHGG) and Contig, Scaffold, Chromosome, and Complete genome (all RefSeq). The genomes in HumGut16S were categorized into the mode category for all the observations in the genome. This means the category occurs most frequently within the genome. One weakness with this is that when finding the mode genome category, the source is not considered. This means that genomes with, for instance, Isolate (a UHGG genome category) as the mode genome category might have RefSeq as the mode source and the other way around.

Figure 2 shows the fraction of genomes matched and not matched by the probes with the different genome categories as mode category. The UHGG categories make a larger fraction of the genomes not matched by the probes than those that are matched. This might indicate that RefSeq has better quality on the 16S sequence and is thus being matched by more probes.

This indication is supported by previous research in HumGut (Hiseni et al., 2021) and similar studies using fecal samples from humans (Meziti et al., 2021). Long-read MAGs might solve this problem.

As mentioned, MAGs in HumGut are short-read MAGs, and all of them are considered to be of high quality (Hiseni et al., 2022). Still, these results, as well as a previous study (Hiseni et al., 2022), indicate that the MAG 16S quality in HumGut might be poorer than the RefSeq 16S quality. This could be because of poorer assembly.

4.2.2 Matches between GA-probe and 16S sequence

Of 4253 genomes, 3134 have at least one probe matching. This means that nearly three-fourths of HumGut16S are discovered by diagnostic tools from GA. As shown in Figure 3, most of the matched genomes have one target, but up to 35 probes match one genome. This is assumed not to be the case and might have something to do with poorly assembled UHGG-genomes.

As explained earlier, the number of genomes is the number of clusters, a group of genomes that are 97.5 % similar across the genome. For genomes with this high similarity, it is expected that the 16S is more similar within the cluster than the 16S sequences in general because the 16S sequence is highly conservative (Langille et al., 2013). Thus, the variation in 16S within one cluster is anticipated to be small.

With this in mind, the number of probes matching shown in Figure 3 and Figure 4 is the sum of probes matching at least one cluster member. This means that three probes matching can mean that the three probes match all the cluster members or that all members are matched by one probe except for one that is matched by two different probes. Because the variance in the 16S sequence is expected to be minor, the cluster members are expected to mainly be matched similarly. Still, one cluster has up to 35 matches in Figure 3. The probes match differently, as shown in Figure 5, and no cluster is expected to be matched by so many probes.

Thus, the 16S sequence variance might be greater than expected. When the number of probes is based on probes that match at least one cluster member, poorly assembled cluster members can affect the analysis. To avoid this, it was attempted only to consider a probe to match the

cluster if at least half of the members were matched. However, some probes are highly narrow-matching, exploit slight variation in 16S-sequences, and might not truly match the whole cluster. Setting such limits for when a cluster is considered matched would exclude these probes.

In addition to analyzing the number of probes matching the genomes for HumGut16S, the same analysis was performed for the subset of HumGut16S containing RefSeq-genomes only. RefSeq genomes are curated by NCBI and are previously shown to be significantly better quality than MAGs (Hiseni et al., 2022), which make out most of the UHGG genomes in HumGut. Using RefSeq-genomes only would therefore exclude the genomes with the assumed lowest quality.

This analysis is shown in Figure 4, which shows the number of probes matching the 2936 HumGut16S RefSeq genomes. Compared to Figure 3, showing the same for all HumGut16S genomes, both figures show that one probe matching the genome is the most common. When the UHGG observations were removed, the maximum number of probes targeting one genome was 10, compared to 35 with the UHGG observations. This indicates that UHGG has a greater variance in the 16S sequences than RefSeq. As explained earlier, the number of probes is the number of probes that matches at least one cluster member. The most matched single genomes have six matches for UHGG and five matches for RefSeq. All genome types but MAG has five matches on one genome as max. MAG has three genomes matched six times that belong to three different clusters.

Genera names were extracted from the HumGut-table and were sometimes inaccurate. *Firmicutes* were received as the genus name for 263 single genomes in HumGut16S, which includes 60 different clusters. Their NCBI organism name in HumGut was *Firmicutes* bacterium, most with numbers after, for instance, *Firmicutes bacterium CAG:884*. Searching these up in NCBI did not gain known genera. To have a genus for these observations, *Firmicutes* is used even though it is not a genus.

Phylum are the highest taxonomic rank used in this thesis, and HumGut16S contains 15 phyla. Of these, 11 is matched by the probes. The phyla *Elusimicrobia*, *Candidatus Saccharibacteria*, *Synergistetes* and *Lentisphaerae* has no genome in HumGut16S that has a

match against any probe. These four phyla are all among the six phyla with fewest HumGut-genomes. Of 4253 genomes, these four phyla in total make 20.

The HumGut16S subset containing only RefSeq-genomes contains 300 genera. Of these, 81 have no genome that is matched by the probes. Of 2936 RefSeq-genomes, these genera are 315. Twenty-nine of these are found in under 1 % of the sequenced human guts. HumGut contains only genomes found in the healthy human gut. Whether it matters that the GA-dysbiosis map does not match these genera depends on their biological function. Two of these genera are *Brevundimonas* (found in almost three percent of the guts) which might be a pathogen (Liu et al., 2021), and *Adlercreutzia* (found in five percent of the guts), which might have a positive health impact (Goris et al., 2021). The latter has previously been shown to be related to dysbiosis (Shaw et al., 2016).

Even though no genomes in HumGut16S belonging to these phyla and genera are matched by the GA-map, there might be HumGut-genomes without available 16S sequences that are matched.

4.2.3 GA-map probes

Figure 5 shows how many genomes, species, and genera the different probe matches. Due to the previous results on genome quality, this is done for RefSeq-genomes only. IG0053 is the probe that matches with most genomes, while IG0023 matches most species and genera. The figure clearly shows that some probes are narrow and others match broader.

Table 3.1 shows that IG0053 matches 874 genomes. In total, 2246 genomes are matched at least once (shown in Figure 4). Thus, IG0053 matches above one-third of the matched genomes. IG0023 and IG0053 are both named after the *Bacillota*-phylum in Table 3.1.

Figure 5 shows that even though IG0053 matches many genomes, only seven probes match above 250 genomes. Of 48 probes, eight have less than 75 % of their matched genomes within the same genera. This supports the expectation that no genome should have 35 matches. This is also supported by the fact that no single genome is matched by more than six probes and that the 16S sequences are expected to be similar. As shown in Table 3.1, two probes are not

matching any RefSeq genomes. It was considered to use RefSeq-genomes only for this part, but because RefSeq misses many genomes and these two clusters, this was not ideal.

4.3 Functional profiling

Functional profiling seems to have good quality for all genomes categories. The GA-map misses some smaller regions of functional space. This is especially the case for HumGut, because the GA-map can only be analyzed for HumGut16S. MAGs might be functionally different from other genomes.

4.3.1 Building functional profiles

4.3.1.1 Gene prediction

The software Prodigal was used to predict coding genes. As stated in the introduction, gene prediction tools like Prodigal overpredicts genes. Prodigal gives a score, and the higher the score is, the more likely it is that the gene is real. Figure 6 shows three random DNA-"genomes" containing no genes and three real genomes. As expected, Prodigal predicts genes in both the randomized DNA and the genomes, but the predicted genes in the genomes generally have a higher score.

It was decided to have a score limit of 30. This is shown with a red line in Figure 6. The figure shows that nearly all the predicted genes in the random DNA are below the limit, and so are many predicted genes in the real genomes. Some of these genes are likely to be real genes. However, using too low a score would produce false positive genes. The figure shows that a lower score would have led to the inclusion of a higher number of genes from the random DNA, and thus also other false genes

4.3.1.2 KEGG ortholog

Figures 7 and 8 show how the KEGG orthologs (KOs) are distributed across HumGut16S (7) and HumGut (8). Both figures show that most KOs are found in a few genomes. No KO was found in all genomes, neither for HumGut nor for HumGut16S. This is surprising because some KOs are expected to be core categories and needed for the microorganisms to be alive.

Furthermore, all these genomes are found in the same environment, a healthy human gut, and it is expected that there are some similarities.

There might be several reasons why no core categories were found. As explained earlier, the chosen prodigal score limit is expected to have filtered out some genes. Thus, core genes might have been filtered out in some genes. All the genomes in HumGut are expected to have high quality Hiseni et al., 2022). Still, there are indications from both previous research (Hiseni et al., 2022) and this thesis that MAGs might have poorer quality than expected. If the assembly quality is poor, there might be missing regions where the core genes are found. It is also possible that there are core genes with multiple functions. As stated in the methods, the sensitivity of Diamond homology searching has been improved (Buchfink et al., 2021). Still, some homologs might not have been identified due to differences in sequence divergence or alignment quality.

4.3.2 Analyze of the functional profiles

4.3.2.1 Using functional profiles to separate genomes into phyla

Figure 9 and 10 clearly shows that genomes of the same phyla group together in functional space. This indicates that the different phyla are separated in functional space and that PCA can detect this with only two components. Hence, the differences between genome categories in the 16S studies can not be seen here.

Figure 9 shows the PCA plot for HumGut16S. This figure shows that for HumGut16S, most genomes from *Bacillota*, *Proteobacteria,* and *Bacteroidota* can be functionally separated from other genomes by using these two PCs. The other phyla seem to be tighter grouped with some subset of different phyla. Plotting PC3 and PC4 makes some of these phyla, for instance, *Actinobacteria,* clearer separated.

Figure 10 shows a somewhat better separation for HumGut. As in Figure 9, *Bacillota* is partially separated from other genomes and partly similar to different phyla. At least most *Bacteroidota* is clearly separated from other genomes. In this “sky” of *Bacteroidota*, one genome belonging to *Synergistetes* can be seen. *Proteobacteria* seems to have genomes

clearly separated from others down in the left corner, but it also contains genomes more similar to others. Some subgroups of other phyla can be seen.

While Figure 9 shows 24.37 % of the detected variance in HumGut16S, Figure 10 explains only 13.28 % of the variance detected in HumGut. Still, the separation seems to be somewhat better. This indicates that the different phyla in fact are separated in functional space and that PCA can detect this with only two components. That phyla are similar in function might seem obvious, but previous studies have shown that taxonomy and function are not always tightly linked (Burke et al., 2011)

4.3.2.2 How HumGut16S is distributed in HumGut

Figure 11 shows no clear separation between genomes that are and are not in HumGut16S. Thus, it seems like HumGut16S can be used as a representative for HumGut. This does not mean that the distribution is perfect. For instance, there appear to be more HumGut16S-genomes in the lower left corner than in the upper left.

4.3.2.3 Functional profiling and match of a GA-probe

Figure 12 shows that IG0023 matches mostly a subset of the functional space of HumGut16S. The probe does not fully match this functional space, and there are matched genomes in other functional regions. Most matched genome is still grouped in one area of the PCA plot, which means that it seems like probes matching and functional profile is related.

That the separation is not perfect could indicate several things. This analysis compares 16S-sequences, which match the probes, and functional profiles. Poor quality on one of them would therefore affect these analyses. Twenty-two matched outliers were identified with a PC1 above -2 and PC2 above 5. All these genomes were MAGs.

Along PC2, there are genomes both with and without match from IG0023 mixed. While some of this may be explained by poor 16S quality or even poor functional profiling, it seems likely that this probe does not match all of this functional space. The probes are not aiming for functional profiles. Thus, IG0023 might truly have openings in the functional area of its matched genomes.

4.3.2.4 Variation in functional profiling between different genome categories

In general, MAGs seem to span out the functional space. This means that the functional profiles seem similar between the different genome categories. Thus, MAGs do not seem to differ from other genome types functionally. This indicates that there are no, or only small, quality issues with the functional profiling of MAGs.

HumGut consists of genomes from two sources (RefSeq and UHGG) and, in total, six genome categories (complete genome, chromosome, scaffold, and contig (all RefSeq) and MAG and Isolate (both UHGG). RefSeq is previously shown to have better genome quality (Hiseni et al., 2021). If the genome categories are functionally different, it would likely be because of this difference in quality.

Figure 13 shows some tendency of grouping between genome categories in HumGut16S, but this is vague. The upper right corner seems to be primarily MAGs (although some Refseq and Isolates are there as well), and a small group of RefSeq-genomes at PC1 around 0 and PC2 around -6 is not visible in the UHGG panel. Figure 14, showing the same for HumGut, also shows similar vague trends.

One reason for the trends to be vague is that many other factors affect the genome's functional profile. One of these factors is taxonomy. To avoid this, a similar figure including only data for one taxon was made. Figure 15 shows a PCA plot for the phylum with the most genomes, *Bacillota*. This figure shows the same as the previous two: there is some weak grouping tendency, but not a perfect one.

To investigate if the vague trend exists or if the pattern is due to other factors, a PLS plot was made (Figure 16). PLS maximizes the variance between the two sources (Wold et al., 2001), meaning that a trend should be visible here. This plot shows a greater tendency to separation between UHGG and RefSeq. This might indicate that it is possible to find patterns in the functional profiles that separate RefSeq-genomes from UHGG and that there might be a systematic bias.

To further narrow down expected variation from different genomes, Figure 17 shows the same for one genus, *Streptococcus* only. *Streptococcus* is not the most dominant *Bacillota*-genera in terms of genomes, but the most dominant ones have very few RefSeq-genomes. This strengthens the hypothesis that there might be a systematic bias. However, since the genus has relatively few observations, it might not be representative.

Accuracy for the two PLS plots is calculated from the same data used to make the model. Consequently, the obtained accuracy is falsely high because this data is used to establish the separation. In this case, PLS was used to test whether a separation is possible. Ideally, a subset of the observations should have been withheld so accuracy could have been calculated based on unseen data. However, for *Streptococcus*, as many observations as possible were needed to make a robust model. For *Bacillota*, the functional differences within the phylum make it difficult to remove observations. Filtering out data to calculate accuracy could have impact on the result. Nevertheless, the obtained accuracy showed that a separation is possible for these data, and that there are some patterns that distinguish between MAGs and other genome types.

As stated, PLS maximizes the variance between the two sources. The *Bacillota*-dataset contains 19567 observations, and *Streptococcus* contains 1088 and 7925 functional categories. This means that for *Streptococcus*, there are many more categories than observations. Because machine learning as PLS combines data in different ways (Obermeyer & Emanuel, 2016; Gourvénec et al., 2003), a pattern found is natural when there are more categories than observations.

The correlation was calculated to investigate whether it is a pattern between functional profile and genome category. Of the 7925 KOs, the correlation could not be calculated for 4999 because these were equal. Thus, these KOs were found in all *Streptococcus* genomes or not found in any. Two were found in all, and 4997 were not found in any. Given that all these genomes belong to the same genera, it is surprising that only two were found in all of these genomes. There are many other KOs found in most of these genomes.

Two functional categories had similar correlations (0.74 and 0.75), which was also the highest correlation found in this study. One of these categories was K12295, which, as explained in the results in section 3.3.2.4, is a two-component histidine kinase response. The other one was

K12294, which is a two-component histidine sensor. It is believed that such two-component histidine kinase sensors recognize environmental signals to influence the activity of a transcription factor they are mated to (Szurmant et al., 2008). There are genomes where one is found and not the other. Still, K12294 and K12295 may be coupled. This would explain why both the functional categories with the highest correlation are related to a two-component histidine system.

These two functional profiles are found in most RefSeq-*Streptococcus* genomes and a few UHGG-*Streptococcus* genomes. It might be random reasons for them to be more frequent in RefSeq than in UHGG. This is especially relevant because the correlation is based on relatively few observations (583 RefSeq and 505 UHGG). Another reason for these two to be the most associated with source could be that the two-component histidine kinase system is complex. Like the 16S sequence, these genome segments are conservated and may have several copies (Eguchi et al., 2017). Thus, assembling these regions may be difficult, which could be why they are not found in UHGG-genomes (mostly MAGs) if these regions are present there.

4.3.2.5 Clustering with K-means

The discovered patterns were studied further using K-means. PCA plot uses two dimensions, while K-means do not reduce the dimensions. This means that K-means can use more of the variation found.

4.3.2.5.1 Clustering HumGut16S with K-means

Of the 43 K-means clusters HumGut16S were clustered into, the GA-map matched 41 to some degree. This means that most of the functional space of HumGut16S matches the map. Figure 18 also shows that both low and high PC1 and PC2 have some matched clusters. Thus, the two PCs explaining most of the variation found do not separate the clusters based on the degree of the match against the GA-map.

On the one hand, the two non-matched clusters are not located on the far sides of the PCA plot in Figure. This indicates that these clusters are not “extreme” and do not exhibit extreme characteristics compared to clusters matched by the probes. Especially the smallest non-

matched cluster is located close to a matched cluster. This means that the non-matched and matched cluster share functional patterns and that these two clusters are not extreme in terms of functional characteristics. On the other hand, the two non-matched clusters are positioned diagonally across from each other in the figure, with an empty gap between them. This could indicate that there is a region that is not matched.

There are 60 functional categories that no matched genome contains. Except for two of them, they are all found in a genome in a cluster that is at least partially matched. This means that even though the functional category is not found in any matched genome, a genome with a similar functional profile as a genome with the category is matched. Thus, these categories are "indirectly matched" by the GA-map.

4.3.2.5.2 Clustering HumGut with K-means

The clustering of HumGut showed larger, darker regions (Figure 19). This is expected because matching with the probes can only be analyzed for genomes in HumGut16S. Thus, the match rate is lower and dark color is more natural. A few groups in this plot stick out. Most genomes with low PC1 are not matched by the probes. These genomes have *Bacillota* as a phylum. *Bacillota* is the phylum with the most genomes in HumGut, as shown in Figure 1. As seen in Figure 10, it is spread over a large part of the functional space, and some genomes are located at the far end. Thus, it is natural that some of these are along the outer edge of this plot.

Another distinct group is located isolated at PC2 above 5. These genomes are *Bacteroidota*, except for one that is *Synergistetes*. If Figure 10 is studied closely, one can see that within the orange "island" of *Bacteroidota*, there is one purple dot. This *Synergistetes*-genome is a MAG that is not in HumGut16S. Thus, this group in Figure 19 probably represents parts of the Bacteroidota-piece of Figure 10.

Nine other *Bacteroidota*-genomes are represented in the clusters in the lower right corner. These clusters are the only distinct group with a high fraction of matched genomes. The other genomes in these clusters are *Proteobacteria*. These genomes are likely to be those shown with orange color in the middle of *Proteobacteria* in the lower left corner of Figure 10. The plot has ten *Bacteroidota*-genomes located in the middle of *Proteobacteria* with PC1 lower

than -10 in Figure 10. Of these, eight are RefSeq-genomes in HumGut16S that are matched by the probes.

Even though some functional areas have a low matched rate, all the groups of clusters have matched genomes. Thus, even though there are clusters without match, there is no large region consisting of several clusters that are never matched by any probe.

It was randomly decided to have around 100 genomes on average in each cluster. In this project, it was decided to have comparable sizes of the K-means clusters for HumGut and its subset HumGut16S to make comparisons more accurate.

4.4 Concluding remarks and further perspective

The first aim of this thesis was to get a taxonomic overview of HumGut and find out how the GA-map covers the taxonomic classifications. GA-map matches can only be measured for the part of HumGut in HumGut16S. Thus, probe matches for most of HumGut remain unknown. Of 16 phyla in HumGut, 11 have at least one genome in HumGut16S that is matched by the GA-map. One phylum is absent in HumGut16S, while four are present with no match. This indicates that the GA-map do not span over the whole taxonomic specter of HumGut. However, the phyla that are not matched are among the phyla with least HumGut-genomes and they also make a small fraction of HumGut16S. These phyla might have HumGut-genomes that are matched. The findings show that the map covers most phyla.

The second aim was to get a functional overview of HumGut and find out how the GA-map spans over the functional space of HumGut. This was done using PCA and K-means clustering. HumGut16S seems to span well over the functional area of HumGut. Some K-means clusters were not matched by the map, indicating that there are functional regions that are missed. The functional profiles missed by the map do not seem to be extreme. In this thesis, probes and functional profiles are mainly studied separately. In the future, it might be interesting to combine this and find which probes match which functional profiles.

The last aim was to evaluate the quality of the genomes in HumGut. This was done by studying GA-map coverage for the different genome categories and with PCA and PLS. The finding suggests that the 16S-sequences from Metagenome Assembled Genomes (MAGs)

might have poorer quality than other 16S-sequences in HumGut. It seems like the quality of the whole genome is better, giving better results on functional profiling. This indicates that the difference found in 16S-sequences is due to these sequences and not that the genomes are different. There are, however, also findings from functional profiling indicating that there could be differences between the functional profiles of MAGs and other genomes.

Given the results from analyzing the quality of genomes in HumGut, this field should be focused more on in the future. Results from this study show that MAGs might have poorer 16S quality than others, but it has yet to be investigated whether some patterns can be found to decide which MAGs have poorer quality. Results from functional profiling might indicate that some functional categories are seen in fewer MAGs than expected. Finding patterns on which categories this applies might be useful for later studies.

Bibliography

Akehurst, R. L., Brazier, J. E., Mathers, N., O'Keefe, C., Kaltenthaler, E., Morgan, A., Platts, M., & Walters, S. J. (2002). Health-related quality of life and cost impact of irritable bowel syndrome in a UK primary care setting. *PharmacoEconomics*, *20*(7), 455–462. https://doi.org/10.2165/00019053-200220070-00003

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., Pollard, K. S., Sakharova, E., Parks, D. H., Hugenholtz, P., Segata, N., Kyrpides, N. C., & Finn, R. D. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature biotechnology*, *39*(1), 105–114. https://doi.org/10.1038/s41587-020-0603-3

Almutairy, M., & Torng, E. (2017). The effects of sampling on the efficiency and accuracy of k-mer indexes: Theoretical and empirical comparisons using the human genome. *PloS one*, *12*(7), e0179046. https://doi.org/10.1371/journal.pone.0179046

Althani, A. A., Marei, H. E., Hamdi, W. S., Nasrallah, G. K., El Zowalaty, M. E., Al Khodor, S., Al-Asmakh, M., Abdel-Aziz, H., & Cenciarelli, C. (2016). Human Microbiome and its Association With Health and Diseases. *Journal of cellular physiology*, *231*(8), 1688–1694. https://doi.org/10.1002/jcp.25284

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amaral, F. A., Sachs, D., Costa, V. V., Fagundes, C. T., Cisalpino, D., Cunha, T. M., Ferreira, S. H., Cunha, F. Q., Silva, T. A., Nicoli, J. R., Vieira, L. Q., Souza, D. G., & Teixeira, M. M. (2008). Commensal microbiota is fundamental for the development of inflammatory pain. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(6), 2193–2197. https://doi.org/10.1073/pnas.0711891105

Buchfink, B., Reuter, K., & Drost, H. G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature methods*, *18*(4), 366–368. https://doi.org/10.1038/s41592-021-01101-x

Börnigen, D., Morgan, X. C., Franzosa, E. A., Ren, B., Xavier, R. J., Garrett, W. S., & Huttenhower, C. (2013). Functional profiling of the gut microbiome in disease-associated inflammation. *Genome medicine*, *5*(7), 65. https://doi.org/10.1186/gm469

Casén, C., Vebø, H. C., Sekelja, M., Hegge, F. T., Karlsson, M. K., Ciemniejewska, E., Dzankovic, S., Frøyland, C., Nestestog, R., Engstrand, L., Munkholm, P., Nielsen, O. H., Rogler, G., Simrén, M., Öhman, L., Vatn, M. H., & Rudi, K. (2015). Deviations in human gut microbiota: a novel diagnostic test for determining dysbiosis in patients with IBS or IBD. *Alimentary pharmacology & therapeutics*, *42*(1), 71–83. https://doi.org/10.1111/apt.13236

Dai, D., Zhu, J., Sun, C., Li, M., Liu, J., Wu, S., Ning, K., He, L. J., Zhao, X. M., & Chen, W. H. (2022). GMrepo v2: a curated human gut microbiome database with special focus on

disease markers and cross-dataset comparison. *Nucleic acids research*, *50*(D1), D777–D784. https://doi.org/10.1093/nar/gkab1019

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology*, *72*(7), 5069–5072. https://doi.org/10.1128/AEM.03006-05

Diaz Heijtz, R., Wang, S., Anuar, F., Qian, Y., Björkholm, B., Samuelsson, A., Hibberd, M. L., Forssberg, H., & Pettersson, S. (2011). Normal gut microbiota modulates brain development and behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(7), 3047–3052. https://doi.org/10.1073/pnas.1010529108

Dimonaco, N. J., Aubrey, W., Kenobi, K., Clare, A., & Creevey, C. J. (2022). No one tool to rule them all: prokaryotic gene prediction tool annotations are highly dependent on the organism of study. *Bioinformatics (Oxford, England)*, *38*(5), 1198–1207. https://doi.org/10.1093/bioinformatics/btab827

Dubey, A. K., Uppadhyaya, N., Nilawe, P., Chauhan, N., Kumar, S., Gupta, U. A., & Bhaduri, A. (2018). LogMPIE, pan-India profiling of the human gut microbiome using 16S rRNA sequencing. *Scientific data*, *5*, 180232. https://doi.org/10.1038/sdata.2018.232

Eguchi, Y., Okajima, T., Tochio, N., Inukai, Y., Shimizu, R., Ueda, S., Shinya, S., Kigawa, T., Fukamizo, T., Igarashi, M., & Utsumi, R. (2017). Angucycline antibiotic waldiomycin recognizes common structural motif conserved in bacterial histidine kinases. *The Journal of antibiotics*, *70*(3), 251–258. https://doi.org/10.1038/ja.2016.151

Gehrig, J. L., Portik, D. M., Driscoll, M. D., Jackson, E., Chakraborty, S., Gratalo, D., Ashby, M., & Valladares, R. (2022). Finding the right fit: evaluation of short-read and long-read sequencing approaches to maximize the utility of clinical microbiome data. *Microbial genomics*, *8*(3), 000794. https://doi.org/10.1099/mgen.0.000794

Goris, T., Cuadrat, R. R. C., & Braune, A. (2021). Flavonoid-Modifying Capabilities of the Human Gut Microbiome-An In Silico Study. *Nutrients*, *13*(8), 2688. https://doi.org/10.3390/nu13082688

Gourvénec, S., Fernández Pierna, J.A., Massart, D.L., & Rutledge, D.N. (2003). Chemometrics and Intelligent Laboratory Systems. PLS regression for the selection of variables based on various stability criteria, 68(1-2), 41-51. https://doi.org/10.1016/S0169-7439(03)00086-8

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., … Gene Ontology Consortium (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic acids research*, *32*(Database issue), D258–D261. https://doi.org/10.1093/nar/gkh036

Hartigan, J. A. & Wong, M. A. (1979). A k-means clustering algorithm. *JSTOR: Applied Statistics*, *28*, 100--108.

Hiseni, P., Rudi, K., Wilson, R. C., Hegge, F. T., & Snipen, L. (2021). HumGut: a comprehensive human gut prokaryotic genomes collection filtered by metagenome data. *Microbiome*, *9*(1), 165. https://doi.org/10.1186/s40168-021-01114-w

Hiseni, P., Snipen, L., Wilson, R. C., Furu, K., & Rudi, K. (2022). Questioning the Quality of 16S rRNA Gene Sequences Derived From Human Gut Metagenome-Assembled Genomes. *Frontiers in microbiology*, *12*, 822301. https://doi.org/10.3389/fmicb.2021.822301

Hou, K., Wu, Z. X., Chen, X. Y., Wang, J. Q., Zhang, D., Xiao, C., Zhu, D., Koya, J. B., Wei, L., Li, J., & Chen, Z. S. (2022). Microbiota in health and diseases. *Signal transduction and targeted therapy*, *7*(1), 135. https://doi.org/10.1038/s41392-022-00974-4

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, *11*, 119. https://doi.org/10.1186/1471-2105-11-119

Ichinohe, T., Pang, I. K., Kumamoto, Y., Peaper, D. R., Ho, J. H., Murray, T. S., & Iwasaki, A. (2011). Microbiota regulates immune defense against respiratory tract influenza A virus infection. *Proceedings of the National Academy of Sciences of the United States of America, 108(13)*, 5354–5359. https://doi.org/10.1073/pnas.1019378108

Jin, H., Hu, G., Sun, C., Duan, Y., Zhang, Z., Liu, Z., Zhao, X.-M., Chen, W.-H. (2022). mBodyMap: A curated database for microbes across human body and their associations with health and diseases. *Nucleic Acids Research, 50(D1)*, D808–D816. https://doi.org/10.1093/nar/gkab973

Kanehisa, M., Goto, S., Kawashima, S., & Nakaya, A. (2002). The KEGG databases at GenomeNet. *Nucleic acids research*, *30*(1), 42–46. https://doi.org/10.1093/nar/30.1.42

Kim, G. H., Lee, K., & Shim, J. O. (2023). Gut Bacterial Dysbiosis in Irritable Bowel Syndrome: a Case-Control Study and a Cross-Cohort Analysis Using Publicly Available Data Sets. *Microbiology spectrum*, *11*(1), e0212522. https://doi.org/10.1128/spectrum.02125-22

Lagkouvardos, I., Overmann, J., & Clavel, T. (2017). Cultured microbes represent a substantial fraction of the human and mouse gut microbiota. *Gut microbes*, *8*(5), 493–503. https://doi.org/10.1080/19490976.2017.1320468

Langille, M., Zaneveld, J., Caporaso, J., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology, 31(9)*, 814–821. https://doi.org/10.1038/nbt.2676

Liland, K. H., Mevik, B.-H., & Wehrens, R. (2022). *pls*: Partial Least Squares and Principal Component Regression [Computer software]. *R package version 2.8-1*. Retrieved from https://CRAN.R-project.org/package=pls

Liu, L., Feng, Y., Wei, L., & Zong, Z. (2021). Genome-Based Taxonomy of *Brevundimonas* with Reporting *Brevundimonas huaxiensis* sp. nov. *Microbiology spectrum*, *9*(1), e0011121. https://doi.org/10.1128/Spectrum.00111-21

Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome medicine*, *8*(1), 51. https://doi.org/10.1186/s13073-016-0307-y

Lovell, R. M., & Ford, A. C. (2012). Global prevalence of and risk factors for irritable bowel syndrome: a meta-analysis. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association*, *10*(7), 712–721.e4. https://doi.org/10.1016/j.cgh.2012.02.029

Lukashin, A. V., & Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic acids research*, *26*(4), 1107–1115. https://doi.org/10.1093/nar/26.4.1107

Maguire, F., Jia, B., Gray, K. L., Lau, W. Y. V., Beiko, R. G., & Brinkman, F. S. L. (2020). Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands. *Microbial genomics*, *6*(10), mgen000436. https://doi.org/10.1099/mgen.0.000436

Maidak, B. L., Olsen, G. J., Larsen, N., Overbeek, R., McCaughey, M. J., & Woese, C. R. (1997). The RDP (Ribosomal Database Project). *Nucleic acids research*, *25*(1), 109–111. https://doi.org/10.1093/nar/25.1.109

Meziti, A., Rodriguez-R, L. M., Hatt, J. K., Peña-Gonzalez, A., Levy, K., & Konstantinidis, K. T. (2021). The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample. *Applied and environmental microbiology*, *87*(6), e02593-20. https://doi.org/10.1128/AEM.02593-20

Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S., & Kyrpides, N. C. (2019). New insights from uncultivated genomes of the global human gut microbiome. *Nature*, *568*(7753), 505–510. https://doi.org/10.1038/s41586-019-1058-x

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *The New England journal of medicine*, *375*(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*, 559-572. https://doi.org/10.1080/14786440109462720

Perisin, M., Vetter, M., Gilbert, J., Bergelson, J (2016). 16Stimator: statistical estimation of ribosomal gene copy numbers from draft genome assemblies. *ISME J 10*, 1020–1024. https://doi.org/10.1038/ismej.2015.161

Pollock, J., Glendinning, L., Wisedchanwet, T., & Watson, M. (2018). The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. *Applied and environmental microbiology*, *84*(7), e02627-17. https://doi.org/10.1128/AEM.02627-17

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, *35*(Database issue), D61–D65. https://doi.org/10.1093/nar/gkl842

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*, *41*(Database issue), D590–D596. https://doi.org/10.1093/nar/gks1219

R Development Core Team. (2010). *R:* A language and environment for statistical computing. *In R Foundation for Statistical Computing*. http://www.R-project.org

Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., Costea, P. I., Godneva, A., Kalka, I. N., Bar, N., Shilo, S., Lador, D., Vila, A. V., Zmora, N., Pevsner-Fischer, M., Israeli, D., Kosower, N., Malka, G., Wolf, B. C., Avnit-Sagi, T., … Segal, E. (2018). Environment dominates over host genetics in shaping human gut microbiota. *Nature*, *555*(7695), 210–215. https://doi.org/10.1038/nature25973

Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K., & Narasimhan, G. (2020). So you think you can PLS-DA?. *BMC bioinformatics*, *21*(Suppl 1), 2. https://doi.org/10.1186/s12859-019-3310-7

Salzberg, S. L., Delcher, A. L., Kasif, S., & White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic acids research*, *26*(2), 544–548. https://doi.org/10.1093/nar/26.2.544

Sangwan, N., Xia, F. & Gilbert, J.A. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* **4**, 8 (2016). https://doi.org/10.1186/s40168-016-0154-5

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia and analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Seemann T. (2013) *Barrnap* 0.7.0 [Online]. Available from https://github.com/tseemann/barrnap

Shaw, K. A., Bertha, M., Hofmekler, T., Chopra, P., Vatanen, T., Srivatsa, A., Prince, J., Kumar, A., Sauer, C., Zwick, M. E., Satten, G. A., Kostic, A. D., Mulle, J. G., Xavier, R. J., & Kugathasan, S. (2016). Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. *Genome medicine*, *8*(1), 75. https://doi.org/10.1186/s13073-016-0331-y

Shin, J., Lee, S., Go, M. J., Lee, S. Y., Kim, S. C., Lee, C. H., & Cho, B. K. (2016). Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. *Scientific reports*, *6*, 29681. https://doi.org/10.1038/srep29681

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., & Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics (Oxford, England)*, *23*(10), 1282–1288. https://doi.org/10.1093/bioinformatics/btm098

Szurmant, H., Bu, L., Brooks, C. L., 3rd, & Hoch, J. A. (2008). An essential sensor histidine kinase controlled by transmembrane helix interactions with its auxiliary proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(15), 5891–5896. https://doi.org/10.1073/pnas.0800247105

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., & Koonin, E. V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic acids research*, *29*(1), 22–28. https://doi.org/10.1093/nar/29.1.22

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., Egholm, M., Henrissat, B., Heath, A. C., Knight, R., & Gordon, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature*, *457*(7228), 480–484. https://doi.org/10.1038/nature07540

Vinje, H., Snipen, L., & Liland, K.H. (2016). *Methods for 16S based taxonomic classification of prokaryotes* (Version 1.2) [R package].

Větrovský, T., & Baldrian, P. (2013). The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PloS one*, *8*(2), e57923. https://doi.org/10.1371/journal.pone.0057923

Wemheuer, F., Taylor, J. A., Daniel, R., Johnston, E., Meinicke, P., Thomas, T., & Wemheuer, B. (2020). Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environmental microbiome*, *15*(1), 11. https://doi.org/10.1186/s40793-020-00358-7

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.* https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... & Yutani, H. (2019). *Welcome to the tidyverse. Journal of Open Source Software*, 4(43), 1686. https://doi.org/10.21105/joss.01686

Wilke, C.O (2020). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 1.1.1. https://CRAN.R-project.org/package=cowplot

Wold, S. S., M. Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109-130. https://doi.org/https://doi.org/10.1016/S0169-7439(01)00155-1

Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K.-H., Ludwig, W., Glöckner, F.O., & Roselló-Móra, R. (2008). The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. *Systematic and Applied Microbiology*, *31*(4), 241-250. https://doi.org/10.1016/j.syapm.2008.07.001

Yoon, S. H., Ha, S. M., Lim, J., Kwon, S., & Chun, J. (2017). A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie van Leeuwenhoek*, *110*(10), 1281–1286. https://doi.org/10.1007/s10482-017-0844-4

Yuan, C., Lei, J., Cole, J., & Sun, Y. (2015). Reconstructing 16S rRNA genes in metagenomic data. *Bioinformatics (Oxford, England)*, *31*(12), i35–i43. https://doi.org/10.1093/bioinformatics/btv231

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway