universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Reevaluating known dynamics in the human gut microbiome with a focus on Prevotellaceae and Bacteroidaceae and their connection to Westernized and Non-Westernized lifestyle

verfasst von | submitted by

Stephanie Pejrimovsky BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 875
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Bioinformatik
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Thomas Rattei

Abstract

The widespread adoption of metagenome-assembled genomes (MAGs) has changed the field of microbiome research. While the focus has shifted from individual species toward community-level structures, the inclusion of species genome bins (SGBs) has simultaneously challenged long-standing species definitions. Ongoing taxonomic revisions within the dominant gut microbiome families Prevotellaceae and Bacteroidaceae raise questions about the stability of lifestyle-associated microbiome patterns and previously described co-occurrence dynamics. This thesis utilized updated, taxonomic frameworks to reassess diversity and co-occurrence patterns in the human gut within these families across Westernized and Non-Westernized human populations.

High-quality genomes were extracted from the Unified Human Gastrointestinal Genome (UHGG) collection and reclassified using PhyloPhlAn and GTDB-Tk to generate consistent species-level genome bin (SGB) assignments, including unnamed SGBs. Species diversity was evaluated at the family level to ensure robustness against ongoing taxonomic changes. A total of 1,000 publicly available stool metagenomes were analyzed using MetaPhlAn 4, enabling the inclusion of revised and previously uncharacterized taxa. Differences in species richness between Westernized and Non-Westernized populations were assessed using presence–absence–based diversity metrics. Community structure was further examined using Spearman correlation analysis, co-occurrence network inference, and unsupervised clustering with DBSCAN and Spectral Clustering.

Genome-based taxonomic classification revealed that a substantial fraction of Prevotellaceae diversity remains unnamed (46.55% uSGBs), whereas Bacteroidaceae appears closer to saturation in terms of species discovery (9.52% uSGBs). Despite recent revisions of the *Segatella copri* complex, two additional unnamed SGBs (*Prevotella* SGB1638 and *Prevotella* SGB1635) were identified that may belong to this lineage.

A strong and reproducible contrast between Prevotellaceae-enriched communities in Non-Westernized populations and Bacteroidaceae-enriched communities in Westernized populations ($p < 2.2 \times 10^{-16}$) was confirmed. In addition to this established dichotomy, Lachnospiraceae species emerged as a distinct group that formed a separate cluster. This cluster was characterized by unique co-occurrence patterns and CAZyme profiles. Rather than supporting discrete enterotypes, these findings indicate family-level population gradients shaped by lifestyle, and functional specialization.

Kurzzusammenfassung

Die breite Anwendung von metagenome-assembled genomes (MAGs) hat das Feld der Mikrobiomforschung grundlegend verändert. Während sich der Fokus von einzelnen Spezies zu gemeinschaftlichen Strukturen verschoben hat, stellt die zunehmende Einbeziehung von Species Genome Bins (SGBs) gleichzeitig etablierte Artdefinitionen infrage. Laufende taxonomische Veränderungen innerhalb der Familien Prevotellaceae und Bacteroidaceae werfen Zweifel an lifestyle Mustern sowie bekannter Koexistenzdynamiken auf. In dieser Arbeit wurden aktuelle Herangehensweisen genutzt, um Diversitäts- und Koexistenzmuster innerhalb dieser Familien im menschlichen Darm über westlich geprägte und nicht-westlich geprägte Populationen hinweg neu zu bewerten. High-quality genomes wurden aus der Unified Human Gastrointestinal Genome (UHGG)-Sammlung extrahiert und mithilfe von PhyloPhlAn und GTDB-Tk neu klassifiziert. Dies führt zu konstanten Zuordnungen der Species Genome Bins (SGBs), einschließlich bislang unbenannter SGBs. Insgesamt wurden 1.000 öffentlich verfügbare Stuhlmetagenome mithilfe von MetaPhlAn 4 analysiert, wodurch sowohl umbenannte als auch zuvor nicht charakterisierte Taxa berücksichtigt werden konnten. Unterschiede im Artenreichtum zwischen westlich geprägten und nicht-westlich geprägten Populationen wurden unter Verwendung Präsenz–Absenz-basierter Matrizen untersucht. Die Gemeinschaftsstruktur wurde mittels Spearman-Korrelationsanalysen, Koexistenznetzwerk und Clusterverfahren (DBSCAN und Spectral Clustering) analysiert.

Die Klassifikation zeigte, dass ein erheblicher Anteil der Diversität innerhalb der Prevotellaceae weiterhin unbenannt ist (46,55 % uSGBs), während die Familie Bacteroidaceae weitgehend gesättigt erscheint (9,52 % uSGBs). Trotz kürzlicher Veränderungen des Segatella-copri-Komplexes wurden zwei zusätzliche unbenannte SGBs (Prevotella SGB1638 und Prevotella SGB1635) identifiziert, die möglicherweise dieser Linie zuzuordnen sind.

Ein starker Gegensatz zwischen Prevotellaceae Gemeinschaften in nicht-westlich geprägten Populationen und Bacteroidaceae Gemeinschaften in westlich geprägten Populationen ($p < 2{,}2 \times 10^{-16}$) konnte bestätigt werden. Über diese etablierte Dualität hinaus traten Arten der Familie Lachnospiraceae als eigenständige Gruppe hervor, die einen separaten Cluster bildeten. Dieser Cluster zeichnete sich durch spezifische Koexistenzmuster sowie charakteristische CAZym-Profile aus. Anstelle strenger Enterotypen deuten diese Ergebnisse auf lifestyle geprägte Gradienten hin.

Table of Contents

Tabel of Figures

List of Abbreviations

Abbreviation	Meaning
ANI	average nucleotide identity
CAZymes	Carbohydrate-Active enZYmes
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
EBI	European Bioinformatics Institute
GTDB-Tk	Genome Taxonomy Database Toolkit
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
HMM	Hidden Markov Model
iTOL	Interactive Tree of Life
kSGB	known SGB
LiSC	Life Science Compute Cluster
LPSN	List of Prokaryotic Names with Standing in Nomenclature
MAG	metagenome-assembled genome
MIMAG	Minimum Information about a Metagenome-Assembled Genome
NGS	next-generation sequencing
PAM	partitioning around medoids
PCA	principal component analysis
RED	relative evolutionary divergence
rRNA	ribosomal RNA
ScC	*Segatella copri* complex
SCFA	short-chain fatty acid
SGB	species-level genome bin
UHGG	Unified Human Gastrointestinal Genome
uSGB	unnamed SGB
WMGS	whole-metagenome shotgun sequencing

1 Introduction

1.1 Progression of the research focus from species to community centered

The term "microbiome" is described as the collection of microorganisms that live in a specific environment, mostly bacteria but also viruses, archaea, and eukaryotic microbes (Berg et al., 2020). Microbial communities linked to specific anatomical niches, including the skin, oral cavity, and particularly the gastrointestinal system, are commonly referred to by this term in the context of human biology (Dekaboruah et al., 2020). Historically, the characterization of the human microbiome relied on culturing techniques, which were limited by the fact that most microbes in the human gut are not readily cultivable under standard laboratory conditions (Parfrey et al., 2011). As a result, our knowledge of microbial variety and function was affected.

The development of sequencing tools such as 16S and 18S ribosomal RNA (rRNA) gene sequencing marked a development in the study of microbiomes. By eliminating the requirement for cultivation and allowing researchers to examine microbial populations straight from environmental samples this technique expanded the field (Clarridge, 2004; Case et al., 2007; Garza & Dutilh, 2015). By focusing on an area that shows a beneficial balance between being conserved and still subject to change researchers could quickly create taxonomic profiles of whole microbial ecosystems (Giovannoni et al., 1990).

The structure, diversity, and dynamics of complex microbial communities became a main interest of microbiology (Case et al., 2007; Garza & Dutilh, 2015). Scientists began to focus on how different communities coexist, interact, and influence their host rather than emphasizing the actions of a single species (Turnbaugh et al., 2007). This major development paved the way for an emerging field of microbiome research where the abilities and composition of entire microbial communities could be investigated by using genome-level data from environmental samples (Qin et al., 2010; Garza & Dutilh, 2015).

Microbiome science has undergone a transformation as a result of this shift from a species-centric to a community-level viewpoint. It has made it possible for scientists to examine questions about how microorganisms interact with their human hosts, function as a group, and adjust to environmental changes rather than listing their residents (Sharpton, 2014).

1.2 Genomic insights from environmental samples reveal previously hidden diversity

Tyson et al.'s successful use of environmental shotgun sequencing to rebuild near-complete microbial genomes from a natural acid mine drainage biofilm in 2004 marked an important milestone in the study of microbiomes. For the first time, this work showed that genomes could be extracted straight from complex microbial communities without the need for cultivation. By providing a window into not just who is present but also what they are genetically capable of doing this introduced the practice of genome-resolved metagenomics (Garza & Dutilh, 2015).

The use of single marker genes restricts the resolution particularly for closely related or functionally diverse strains. According to Jovel et al. (2016), this limits direct insights into the functional potential of microbial communities. These limitations gradually caused the field to adopt whole-metagenome shotgun sequencing (WMGS), which sequences almost all genetic material in a sample rather than focusing on marker genes (Sharpton, 2014). This approach captures the full metagenome, which is the collection of genetic material in an environment. Furthermore, according to Marchesi and Ravel (2015), WMGS is capable of capturing all microbial domains. Unlike 16S rRNA sequencing, WMGS offers both high-resolution taxonomic identification and direct insights into functional potential (Sharpton, 2014).

The widespread use of WMGS was made possible by the development and advancement of next-generation sequencing (NGS) technology in the late 2000s. By enhancing sequencing capacity and reducing costs, these improvements have made it possible to perform broad metagenomic research on a variety of environments, including the human gut (Goodwin et al., 2016). Specialized computational methods for assembly, binning, and annotation had continuously evolved to manage the huge amount of generated data (Satam et al., 2023). By the early 2010s, metagenomics was considered a crucial method for studying microbial ecology and human microbiomes (Qin et al., 2010).

1 Introduction

The official introduction of the term "metagenome-assembled genome" (MAG) in 2015 to refer to draft genomes that were reconstructed from metagenomic data highlighted another important moment (Garcia et al., 2015; Hugerth et al., 2015). This method involves assembling short sequencing reads into contigs, grouping those contigs into genome bins, and analyzing genome completeness and contamination (Bowers et al., 2017). MAGs supported researchers to move beyond community-level profiling and carry out genome-centric studies, which revealed novel insights into metabolic capacity and evolutionary histories of uncultured microorganisms (Kim et al., 2024).

The Minimum Information about a Metagenome-Assembled Genome (MIMAG) criteria were presented by Bowers et al. in 2017 and provided official instructions for MAG quality reporting. These guidelines strengthened the reliability of MAGs, improved data exchange, and helped to create standardization across studies. Without a requirement for cultivation, the ability to directly reconstruct genomes from complex environmental samples revealed previously undiscovered microbial diversity (Nayfach et al., 2021).

This genome-centric approach helped uncover some of the hidden diversity, sometimes known as microbial "dark matter," or species that are widespread but have not yet been identified or are not cultivable (Rinke et al., 2013; Hug et al., 2016). Additionally, it helped to understand how different community compositions can occur in niches that are almost identical. One example for this is the human gut microbiome across populations (Pasolli et al., 2019). In summary, MAGs made comparative microbiomics possible in addition to identifying hidden diversity within specific groups. Scientists are now more equipped to analyze microbial lineages, identify ecological and evolutionary patterns, and connect genomic variation to functional aspects because of to this technique (Nayfach et al., 2021; Almeida et al., 2021).

1.3 Metagenomics as a tool to assess microbial diversity across human populations

The ability to reveal the diversity of microbial communities across various human populations is an important aspect of metagenomics (Yatsunenko et al., 2012; Pasolli et al., 2019). The human gut microbiome continues to provide new findings when examined using genome-resolved approaches, even though it is one of the most researched microbial ecosystems (Almeida et al., 2021). By providing both taxonomic and functional information, metagenomics has revealed patterns of variation that were previously difficult to identify through culture-based or marker gene approaches (Jovel et al., 2016).

This change in technology made it possible to analyze genes, metabolic pathways, and microbiological species within a single sample (Franzosa et al., 2015). Therefore, the known differences in composition between populations living in traditional, Non-Westernized contexts and those living in highly industrialized, Westernized countries could be further investigated (Pasolli et al., 2019; Yatsunenko et al., 2012; Segata, 2015).

1.3.1 Defining the Westernized lifestyle

Adopting viewpoints, customs, and cultural norms typically seen in economically developed nations, especially those in Western Europe and North America, is referred to as a "Westernized lifestyle" (Ozdemir et al., 2024). In microbiome studies, samples from people who live in these social settings are usually categorized as "Westernized". This reflects the wider impact of lifestyle variables on human physiology, including the composition of the gut microbiome. Geographical location or a country's level of development are not the only restrictions on the use of the term. In low- or middle-income nations, urban areas may also display Westernization traits such industrialized food systems, excessive antibiotic usage, and less engagement with the environment (Wheaton & Crimmins, 2013; Vinogradova et al., 2024). Therefore, "Westernized" is more a reference to lifestyle choices than to political or economic categorization.

1 Introduction

Westernized dietary patterns are characterized by a low intake of fermented foods, fiber, fruits, vegetables, and a high intake of animal fats, added sugars, ultra-processed foods, and red meat (Sonnenburg & Sonnenburg, 2019; Clemente-Suárez et al., 2023). This nutritional profile contrasts highly with traditional or rural societies, where plant-based diets high in fiber are more prevalent (De Filippo et al., 2010). Because dietary intake has a direct impact on the nutrients that gut bacteria can utilize, these variations have a direct influence on microbial diversity (Wu et al., 2011).

A gut microbiome dominated by bacteria species from the genus *Bacteroides* is frequently associated with a Westernized lifestyle (Wu et al., 2011). These bacteria have a variety of enzymes which makes them capable of breaking down glycans that are both host-derived and associated with the diet. According to David et al. (2013), this makes them ideal for diets heavy in animal proteins and saturated fats.

Gut communities in Westernized populations typically have higher concentrations of specific Carbohydrate-Active enZYmes (CAZymes), especially those that break down simple sugars and mucin (Sonnenburg & Sonnenburg, 2019). The reduced consumption of complex plant polysaccharides and the greater availability of readily digestible carbohydrates in typical Western diets are reflected in this enzymatic profile. Therefore, energy extraction from meals high in fat and protein may be given priority by the Western gut microbiome (Sonnenburg et al., 2016).

Changes in the composition of the microbiome may also be caused by factors that are linked to Westernization, in combination with nutrition. These include sedentary behavior, a greater dependence on technology and digital media, a decrease in exposure to animals and soil bacteria as a result of urban living (Parajuli et al., 2018), and an increase in the usage of antibiotics and medications (Lee et al., 2023). According to Ozdemir et al. (2024), the perception of personal hygiene and cultural practices can also be influenced by educational and social norms. These norms frequently result in a decrease in traditional ways of living, an increase in individualism, and a decrease in interaction with animals and natural settings. When combined, these elements produce an ecological context that is quite distinct from Non-Westernized environments.

1.3.2 Defining the Non-Westernized lifestyle

The term "Non-Westernized" refers to a lifestyle that is distinct from the standards of industrialized western countries. This classification is commonly used to describe populations in geographic regions where rural living, agrarian economies, and traditional cultural practices are still more prominent (Yatsunenko et al., 2012; Martínez et al., 2015), such as South and Southeast Asia (Pingali, 2007; Lipoeto et al., 2012), parts of the Middle East (Naveed et al., 2022), Africa (Afolayan et al., 2019), and Latin America (Parra-Soto et al., 2025; Wang et al., 2021). These lifestyles are frequently, but not necessarily, associated with handmade manufacturing, cattle herding, smaller-scale farming, and close ties to local environmental structures (Moritz et al., 2025).

In general, Non-Westernized diets are rich in dietary fiber and plant-based foods. They tend to favor more fruits, vegetables, whole grains, and legumes rather than processed food, added sugar, and animal fats (Landry & Ward, 2024; De Filippo et al., 2010). Cultural and religious practices, regional farming methods, community-based food preparation, seasonal cycles, and local biodiversity are all closely connected to these dietary habits (Martínez et al., 2015; Rampelli et al., 2015).

Gut microbiomes from people who live Non-Westernized lifestyles are frequently enriched in Prevotellaceae species, especially *Prevotella copri*, in contrast to Westernized samples (De Filippo et al., 2010; Yatsunenko et al., 2012). With its exceptional capacity to break down a broad range of complex polysaccharides originating from whole grains, legumes, and other plant sources, this genus is closely linked to plant-based, high-fiber diets (Aakko et al., 2020).

Studies using metagenomic and enzymatic profiling have repeatedly demonstrated that Non-Westernized microbiomes have a greater variety of CAZymes, particularly those that target resistant starches, hemicelluloses, and pectins (Tett et al., 2019; Wu et al., 2011). According to Chen et al. (2017), this enzymatic ability supports a microbial community that is oriented toward the fermentation of indigestible fibers into useful short-chain fatty acids (SCFAs). For example, the SCFAs butyrate and propionate, which are needed for keeping the integrity of the gut barrier and managing immune responses. The strong evolutionary link between nutrition, environment, and microbial function within traditional lifestyles is highlighted by the abundance of *Prevotella copri* and fiber-degrading CAZymes in these communities (Tannock, 2023).

It's important to understand the variety and nuances that the term "Non-Westernized" includes. It can be used to describe a wide range of cultural settings and living conditions from rural communities in otherwise highly industrialized countries to indigenous populations living in isolated regions (Moritz et al., 2025). Therefore, while utilizing the phrase for broad comparisons in microbiome research, caution should be taken to avoid oversimplifying or generalizing the experiences of diverse individuals (Núñez Casal, 2024).

1.3.3 Sampling bias toward Westernized populations in public microbiome databases

One significant drawback in the research of the human microbiome is the overrepresentation of members in Westernized societies in public sequencing databases (Kim et al., 2024; Almeida et al., 2020). Economic, practical, and historical issues all contribute to this sampling bias. Wealthy, industrialized nations with widespread access to next-generation sequencing methods, financing, and scientific infrastructure were the birthplaces of early microbiome research (Foxx et al., 2021). Consequently, a large number of noteworthy microbiome datasets and reference genome collections were constructed mostly on Westernized cohorts, which are characterized by frequent use of antibiotics, low-fiber diets, and urban living (Porras & Brito, 2019; Kim et al., 2024).

In addition to this historical context, gathering samples from Non-Westernized communities includes major logistical and legal challenges. Data collection can be complicated due to isolated locations, a lack of local sequencing facilities, and the need for ethical approval across several institutions or nations (Foxx et al., 2021). Furthermore, it is often harder to provide high-quality data from traditional or rural communities due to issues with sample transportation, preservation without a cold chain, and obtaining adequate metadata (Bensch et al., 2022). On the other hand, samples from Westernized people are simpler to collect, standardize, and sequence in a lab setting. These samples are frequently taken from urban hospital cohorts or pre-existing biobanks (Abdill et al., 2022).

The reference datasets used by the majority of taxonomy classifiers and genome catalogs have been influenced by this disparity in data. For instance, Western samples made up the majority of large-scale resources such as the Human Microbiome Project and early iterations of genome databases like RefSeq (Arif et al., 2025). As a result, analyses based on these references may underrepresent or misclassify taxa that are common in Non-Westernized individuals or entirely miss organisms unique to them (Carter et al., 2023).

A move toward more inclusive sample techniques has been sparked in recent years by the decreasing cost of sequencing and a greater understanding of global health imbalances (Foxx et al., 2021). To increase the phylogenetic and functional range of human-associated microbial diversity, international projects and partnerships have been formed to sequencing microbiomes from historically underrepresented groups (Nayfach et al., 2020; Pasolli et al., 2019). However, the enduring impact of sampling bias keeps limiting our capacity to draw universally relevant conclusions about human–microbe interactions and is a key distracting factor in comparative microbiome studies (Abdill et al., 2022; Arif et al., 2025).

1.4 Utilization of unnamed Species Genome Bins (uSGBs) to maximize the potential of metagenomic data

Alongside the increasing availability of metagenomic data from Non-Westernized populations, recent methodological innovations have expanded our capacity to extract novel insights from existing datasets (Almeida et al., 2020). One of the most influential developments in this context is the introduction of species-level genome bins (SGBs), a classification framework that organizes the vast amount of genomic information derived from microbial communities (Pasolli et al., 2019).

SGBs allow researchers to group genomes recovered from metagenomic data at an approximate species level, even when the organisms have no formal taxonomic classification or have never been successfully cultivated (Blanco-Míguez et al., 2023). This approach was proven particularly valuable for studying complex microbial ecosystems such as the human gut microbiome, where a large proportion of the microbial diversity remains unknown or insufficiently characterized (Pasolli et al., 2019; Almeida et al., 2020).

Metagenome-assembled genomes (MAGs), which are created by computationally assembling and binning environmental sequencing data, are commonly used to build SGBs (Almeida et al., 2020). In order to create species-level groupings, these bins are compared using metrics like average nucleotide identity (ANI), often with a threshold of approximately 95 percent (Pasolli et al., 2019). Furthermore, SGBs are divided into two categories according to their characterization status. While known SGBs (kSGBs) correspond to previously characterized genomes found in reference databases, unnamed SGBs (uSGBs) show no close resemblance to any identified species (Blanco-Míguez et al., 2023).

An important step in the research field of the microbiomes was made with the introduction and systematic cataloging of uSGBs. It became known that many of the microbial genomes found in the human gut could not be connected to previously identified organisms as sequencing methods advanced and more extensive metagenomic studies were conducted (Li et al., 2024). These uSGBs make up a substantial amount of what is sometimes called "microbial dark matter," which is made up of lineages that are either absent from established taxonomic frameworks or not accessible through traditional culture (Alanzi, 2024; Li et al., 2024). Some uSGBs may belong to established taxonomic groupings but are sufficiently different to be undetectable with existing categorization techniques, while others most likely represent completely new taxa (Blanco-Míguez et al., 2023).

The ability to enable more focused and informed cultivation activities is one of the useful benefits of uSGBs. Researchers may predict specific phenotypic features, such metabolic pathways or environmental tolerances, by reconstructing partial or full genomes from metagenomic data (Liu et al, 2022). By using this knowledge to inform the creation of selective growth conditions, it will be more likely that previously uncultured species will be isolated from complex microbial ecosystems (Salam et al., 2020).

Microbiome science has advanced beyond the constraints of conventional taxonomy and cultivation thanks to the introduction of SGBs, especially uSGBs (Blanco-Míguez et al., 2023). Without depending on pre-existing reference genomes, this genome-based approach enables more thorough and precise characterisation of microbial communities (Pasolli et al., 2019). Because of this, uSGBs are becoming important assets for researching biogeographic patterns, population structure, host specialization, and microbial evolution (Lemos et al., 2021). Additionally, they have made it possible for researchers to investigate the functions that uncharacterized bacteria play in both health and disease, as well as their functional potential (Alanzi, 2024).

1.5 The Unified Human Gut Genome (UHGG) Collection as an example of hidden diversity revealed through metagenomics

A compelling example of how genome-resolved metagenomics has revolutionized our knowledge of the human gut microbiome is the Unified Human Gastrointestinal Genome (UHGG) collection. Although the theoretical potential of metagenome-assembled genomes (MAGs) and species-level genome bins (SGBs) has been covered in earlier chapters, the UHGG collection provides an unparalleled scale of practical demonstration of these ideas (Almeida et al., 2020).

The UHGG collection, which was published by Almeida et al. in 2020, was created by combining genome sequences from eight important datasets, which included metagenome-assembled genomes (MAGs) and cultivated isolates. In total over 286,000 genomes were included which resulted in 204,938 nonredundant genomes following deduplication and quality filtering. With a 95% average nucleotide identity criteria, these were grouped into 4,644 species-level genomic bins. The vast amount of uncultured microbial diversity that remains in the human gut microbiome is highlighted by the fact that 71% of these species lacked a cultivated representative.

The significance of the UHGG collection comes not just from its size but also in its capacity to bring attention to microbial diversity that had previously been widely underestimated. The majority of the uSGBs found in this collection are from uncultivated or previously undiscovered lineages (Almeida et al., 2020). Several of them contain major functional characteristics, including genes related to antibiotic resistance, secondary metabolite production, or carbohydrate metabolism. Older reference genomes, which were mostly constructed from cultivated isolates, lacked these characteristics (Arif et al., 2025). Therefore, the UHGG collection broadens the gut microbiome's functional landscape in addition to its taxonomic resolution.

Additionally, the UHGG collection demonstrated the differences in microbial diversity among human populations. It was discovered that people from Non-Westernized backgrounds had higher prevalences of several uSGBs observed in the collection. According to Abdill et al. (2025), this supports the notion that large pieces of gut microbial diversity have been missed by earlier microbiome research that was primarily based on Western samples. This means that the collection functions as a genetic resource as well as a corrective for the biases that have influenced microbiome research in the past.

Fundamentally, the UHGG collection serves as an example of how to reanalyse current datasets using more comprehensive sampling techniques and updated methods. By using cutting-edge assembly and binning techniques on previously available metagenomic data, the study discovered thousands of unique microbial species without obtaining a single new sample (Almeida et al., 2020). This scalable and affordable method demonstrates how the discipline might develop further alongside the increasing accessibility of sequencing.

1.6 Revisiting prior assumptions: from Prevotella copri to the Prevotellaceae family

For many years, the genus *Prevotella*, particularly *Prevotella copri*, was portrayed as one half of an interesting dualism in the human gut microbiome. Individual gut communities are often described as "*Prevotella*-dominated" as opposed to "*Bacteroides*-dominated" in microbiome studies (Gorvitovskaia et al., 2016). Dietary and lifestyle variables appeared to be the cause of these variances. While *Bacteroides* dominance was usual seen in Westernized populations consuming diets heavy in protein and fat, *Prevotella*, and especially *P. copri*, was often observed among people from Non-Westernized communities eating fiber-rich, plant-based diets (De Filippo et al., 2010; Wu et al., 2011).

Later on, it became clear that this binary approach overlooked a great deal of complexity, especially among *P. copri*. A 2019 study by De Filippis and colleagues that focused on strain-level functional diversity rather than genus-level differences indicated a significant change in how community interactions are approached. The functional ability of *P. copri* strains varied greatly, according to the study, which included metagenomic data from Italian adults who followed vegetarian, vegan, and omnivorous diets.

The strains linked to plant-based diets carried genes that promote the breakdown of complex carbohydrates, including xylans and pectins. The genes for xenobiotic metabolism and branched-chain amino acid biosynthesis, on the other hand, were more abundant in strains derived from omnivorous individuals. The health effects of *P. copri* may vary greatly depending on the strain composition, since some strains linked to Western diets were found to have characteristics that may contribute to insulin resistance and impaired glucose metabolism (De Filippis et al., 2019).

Building on these findings, Tett et al. performed an additional analysis later that year, employing over 1,000 *P. copri* genomes rebuilt from over 6,800 human gut metagenomes from 25 countries. This study demonstrated that *P. copri* is not a single, homogeneous species, but rather a genetically diverse complex made up of four very different clades named A, B, C, and D. These clades diverged by more than 10% in average nucleotide identity, considerably above the threshold generally used to distinguish between species.

These clades could not be distinguished using conventional taxonomic markers because of their highly conserved 16S rRNA sequences. The distinct gene repertoires of each clade, notably in the area of carbohydrate metabolism, provide plausibility to the hypothesis that their evolutionary paths were influenced by long-term dietary selection, particularly fiber intake. Additionally, individuals from Non-Westernized populations frequently shared clades, indicating a long-standing ecological relationship between *P. copri* diversity and dietary trends.

Hitch and colleagues made further improvements to this taxonomy in 2022, moving the *P. copri* complex into the newly formed genus *Segatella*. Blanco-Míguez et al. (2023) used metagenomic reconstruction from over 20,000 samples as well as cultivation-based approaches (44 isolates) to further update the *Segatella copri* complex (ScC) and introduce 13 unique species within it. Nine more species-level clades (E through M) were added to the initial four clades in the new classification. A few of these were host-specific, meaning they only occurred in non-human primates or humans.

Analyzing microbiological patterns at the family level rather than just comparing genera has become more important as taxonomic precision has increased and species placement shifted. This is especially relevant since numerous species have been transferred from the Bacteroidaceae to the Prevotellaceae family, and both groups have seen significant internal replacement. Recent study indicates that analyzing these genera in the context of their larger taxonomic groups is necessary to comprehend their ecological roles and functional capacities, despite the fact that early studies frequently presented them as oppositional (Gorvitovskaia et al., 2016).

This study aims to further address this aspect by generating a comprehensive overview of Bacteroidaceae and Prevotellaceae diversity in the human gut microbiome. Using updated taxonomic frameworks and genome-resolved metagenomic methods, the project aims to utilize SGBs to create a more stable picture of these two prominent human gut families.

1.7 Recontextualizing microbial diversity in the human gut

One of the unique characteristics of microbiome research is the extent of biological complexity involved, particularly in the study of the human gut. The gut microbiome consists of complex, interlinked ecosystems composed of thousands of bacteria species. In addition to being very diverse among individuals, these communities are also sensitive to changes over time inside a single host (Lloyd-Price et al., 2019; Yatsunenko et al., 2012). Many factors, such as immunological activity, the environment, diet, and medication, might affect the microbial makeup (McDonald et al., 2018). This complexity presents an important challenge for researchers trying to understand how the microbiome supports health, responds to disruptions, or increases a person's risk of developing illnesses (Manor et al., 2020).

To handle this complexity, microbiome science has consistently used the grouping concept. It is possible to reduce dimensionality, reveal hidden patterns, and improve the interpretability of biological findings by classifying microbial communities into structured groups. This method spans from early taxonomic classifications to more contemporary methods like enterotype clustering and network-based models of species interactions (Arumugam et al., 2011; Röttjers & Faust, 2018). The large, frequently noisy datasets produced by metagenomic research can be assigned structure using any of these frameworks (Armstrong et al., 2022). The idea of how these frameworks have changed over time, what they tell us about microbial ecology, and how they could be reinterpreted in light of new functional and taxonomic information will all be covered in the parts that follow.

1.7.1 Network-based approaches to characterizing the human gut microbiome

The composition and functional ability of the human gut microbiome are directly related to host health. We can now catalog microbial species much more efficiently due to high-throughput sequencing technology, but they don't reveal much about the relationships and dependencies that make up these communities (Armstrong et al., 2022). Understanding the ecological principles that underlie community assembly, stability, and resilience requires an understanding of how microbial species coexist, compete, or exclude one another (Coyte et al., 2015). Based on patterns of abundance and presence across large populations or lifestyles, co-occurrence and co-exclusion network modeling has become a popular technique for determining possible ecological relationships among microbes in this context (Layeghifard et al., 2017; Armstrong et al., 2022).

These networks serve as ecological models, capturing relationships that might represent fundamental biological processes like commensalism, mutualism, competition, or niche differentiation (Coyte et al., 2015). According to Weiss et al. (2016), a positive correlation, or co-occurrence, between two taxa may indicate syntrophic linkages or comparable environmental preferences. On the other hand, antagonism or competitive exclusion may be indicated by negative correlations, sometimes known as co-exclusion patterns (Weiss et al., 2016). These relationships offer a strong foundation for developing hypotheses, despite the fact that they are strictly correlational and do not suggest causality or physical interaction (Faust & Raes, 2012). Potential keystone species can be found using network models, which can show structured microbial subcommunities and provide information on how microbial associations react to outside disturbances like disease, antibiotic exposure, and dietary changes (Layeghifard et al., 2017).

Connecting microbial structure to nutritional ecology is one particularly interesting use of network analysis. According to comparative studies, diets rich in plants and high in fiber tend to support dense, cooperative microbial networks, which are frequently dominated by taxa that produce short-chain fatty acids (SCFAs) (Nogal et al., 2021). Conversely, more fragmented and competitive networks are associated with Western diets, which are generally heavy in animal protein and fat and low in fiber (Clemente-Suárez et al., 2023). These networks show decreased connectivity and modularity, which could be indicative of microbial communities that are less robust or dysbiotic (Sonnenburg, 2016). These trends have been supported by cross-cultural comparisons between traditional and industrialized populations, showing how nutrition and lifestyle affect the ecological structure of the microbiome in addition to its composition (Wang et al., 2021).

Abundance networks have important methodological and interpretive limitations despite their usefulness. The fact that correlation does not indicate interaction is the most basic problem. Due to their similar responses to common environmental factors, two species may frequently co-occur (Deutschmann et al., 2021). Furthermore, microbial interactions frequently change based on circumstance and are rarely static or binary (Manor et al., 2020; Martínez et al., 2015). Additionally, the majority of network analyses rely on cross-sectional data, which only records a single moment in time and is unable to take temporal dynamics into account (Deutschmann et al., 2023). The fact that network properties, like the discovery of highly connected hub taxa, may be overinterpreted in the absence of additional functional evidence or experimental confirmation is now becoming more well recognized (Faust & Raes, 2012).

Despite these drawbacks, interaction networks have contributed to the reinterpretation of the gut microbiota as an environmentally regulated and structured system (Faust & Raes, 2012). They have been useful in predicting microbial responses to interventions (Coyte et al., 2015), identifying possible biomarkers of health and disease (Röttjers & Faust, 2018; Layeghifard et al., 2017), and directing experimental efforts to characterize microbial functions (Faust & Raes, 2012). Through the lens of networks, microbial populations can be seen as dynamic and interdependent.

In the context of this study, network analysis is an important method for investigating the ecological interactions between Prevotellaceae and Bacteroidaceae members across distinct human populations. This work tries to determine whether these two families adhere to different ecological strategies and whether they occupy mutually exclusive niches within the gut by analyzing patterns of co-occurrence and exclusion at high taxonomic precision. This method adds to a more comprehensive understanding of gut microbial organization and its reliance on environmental and lifestyle factors, in addition to complementing taxonomic and functional profiles (Deutschmann et al., 2021).

1.7.2 Enterotype-based classification of gut microbiomes

How to effectively classify the enormous individual variability in microbial community makeup has been one of the main issues in microbiome research. In an important study published in 2011, Arumugam et al. proposed the idea of "enterotypes," suggesting that human gut microbiomes might be divided into three distinct community types.

Arumugam et al. (2011) used metagenomic data from 39 individuals across different communities by using unsupervised clustering techniques, such as principal component analysis (PCA) and partitioning around medoids (PAM). Three strong clusters, called enterotypes, were identified by their research and each was defined by the dominance of a single bacterial genus. *Bacteroides*-dominated Enterotype 1 was linked to diets high in saturated fats and animal protein. Enterotype 2 was caused by *Prevotella* and was associated with diets high in fiber and carbohydrates. Microbes belonging to Enterotype 3, which was typified by *Ruminococcus* species, had improved mucin-degrading and sugar-fermenting capacities.

It was suggested that the enterotypes were stable ecological configurations that were unaffected by host characteristics including age, gender, and body mass index, as well as geographic origin and sequencing technique. This implied that internal ecological principles, as opposed to external host characteristics, might influence the organization of microbial communities. The enterotype concept's simplicity and potential for clinical use were its primary selling points. In early microbiome research, the enterotype theory consequently received a lot of momentum (Arumugam et al., 2011).

Over time, it became clear that the enterotype model had a number of shortcomings. According to Costea et al. (2018) and Jeffery et al. (2012), one major critique of clustering techniques is that they could artificially define boundaries for what is essentially a continuous distribution of community states. The majority of clustering methods can be sensitive to distance metrics, data processing, and cohort composition, and they require pre-specification of the number of clusters (Koren et al., 2013; Layeghifard et al., 2017). Furthermore, it is frequently impossible to categorize the gut microbiome in a binary or discrete manner because to its ecological complexity. According to Weiss et al. (2016), microbial communities are extremely dynamic and are influenced by constant interactions between the host, the environment, and the microbiota itself. It is challenging to reconcile the diversity introduced by these elements with static categories.

Reproducibility has been another major critique of the enterotype idea. Their universality is called into question by the number and composition of enterotypes reported by investigations using various analytical methods, sequencing depths, or demographic cohorts (Koren et al., 2013; Huse et al., 2012). Knights et al. (2014) argue that rather than being strict categories, classifications of microbiome diversity should be seen as a gradient. In a 2014 editorial, the original authors acknowledged that enterotypes should not be viewed as stable categories but rather as dense areas within a multidimensional continuum. Early interpretations that depicted enterotypes as distinct and stable microbiome types frequently overlooked this complexity (Bork et al., 2014).

Enterotypes now play a different function in microbiome science as a result. They are still used, particularly in global and cross-cultural studies, but only as heuristic tools or conceptual starting points (Costea et al., 2018; Knights et al., 2014). Given these developments, the goal of this work is to reexamine the clustering question in gut microbiome research while accounting for the increased taxonomic precision made possible by genome-resolved metagenomics. Through the examination of microbial co-occurrence patterns and the incorporation of new taxonomic reclassifications, this study investigates whether species form reproducible and ecologically significant clusters.

1.8 Summary and goals for the present study

As discussed in the previous chapters, the development of metagenomic techniques has revolutionized our knowledge of the human gut microbiome by exposing its vast taxonomic and functional diversity. Additionally, they have emphasized the need for more complex, genome-resolved analyses that can take into consideration the differences across a range of human lifestyles especially between Westernized and Non-Westernized ones. Prevotellaceae and Bacteroidaceae stand out in this context as being essential to the architecture of microbial communities.

Therefore, this thesis's three objectives are as follows:

(1) provide an overview of species-level diversity within the Prevotellaceae and Bacteroidaceae families in the human gut microbiome, using updated taxonomic frameworks that include uSGBs

(2) Assess if there are differences in species diversity of Bacteroidaceae and Prevotellaceae in Westernized and Non-Westernized populations when incorporating revised and previously uncharacterized SGBs

(3) Reevaluate co-occurrence dynamics in the human gut microbiome at the family and community level

2 Methods

There are two primary parts to this study's methodological structure. Characterizing the diversity of microorganisms in the human gut, with a focus on the groups Bacteroidaceae and Prevotellaceae, is the first component. High-quality genomes are extracted from the Unified Human Gastrointestinal Genome (UHGG) collection, taxonomic reclassification is performed using genome-resolved methods like PhyloPhlAn and GTDB-Tk, and the resulting species-level genome bins (SGBs) are phylogenetically contextualized.

The second portion looks into patterns of ecological interactions, namely the dynamics of co-occurrence and co-exclusion between species in the two target groups. In order to investigate the structure of microbial communities across various human lifestyles, this section combines metagenomic community profiling, network creation, and unsupervised clustering algorithms. The goal of the investigation is to determine how updated taxonomic frameworks and previously uncharacterized taxa influence gut microbial ecology.

All computational analyses were carried out on the Life Science Compute Cluster (LiSC) of the University of Vienna, utilizing its Unix-based environment along with the JupyterHub and RStudio Server interfaces. All scripts can be found in the supplementary material. The table below provides an overview of all software tools used, including version numbers and corresponding references.

Software	Version	Reference
UHGG	2.0	Almeida et al., 2020
PhyloPhlAn	3.0	Asnicar et al., 2020
GTDB-tk	2.4.0	Chaumeil et al., 2019
iTOL	v7	Letunic & Bork, 2006
curatedMetagenomicData package	1.7.5	Pasolli et al., 2017
MetaPhlAn	4.0	Blanco-Míguez et al., 2023
Scikit-learn	1.6.1	Pedregosa et al., 2011
NumPy	1.26.4	Harris et al., 2020
NetCoMi	1.1.0	(Peschel et al., 2025)
Prokka	1.14.6	(Seemann, 2014)
Run_dbcan	4.1.4	(Zheng et al., 2023)

2.1. Assessing the diversity of Prevotellaceae and Bacteroidaceae in the human gut

As discussed in Chapter 1.6 recent taxonomic revisions have provided a more refined understanding of Prevotellaceae, particularly through the resolution of the *Segatella copri* complex. Alongside, further taxonomic reclassification took place in both Bacteroidaceae and Prevotellaceae. A part of this study wants to continue the work on a systematic, genome-based comparison of both families within a unified analytical framework. This section of the thesis aims to address this by providing a comprehensive overview of species-level diversity in both Prevotellaceae and Bacteroidaceae using genome-resolved methods.

2.1.1 Dataset extraction from the UHGG collection

The Unified Human Gastrointestinal Genome (UHGG) collection (Almeida et al., 2020) is a publicly accessible resource that offers high-quality reference genomes of the human gut microbiome, including isolate genomes and metagenome-assembled genomes (MAGs). This is where the dataset used for this analysis came from. The second version of the UHGG collection is a good starting point for describing microbial diversity at the species level because it includes more than 200,000 nonredundant genomes that have been aggregated from a variety of research and sample origins. Reproducibility is further supported by its extensive coverage and accessibility, which also will allow the results to be confirmed or extended in subsequent research.

Every genome that is part of the UHGG collection is listed in a related metadata file that is provided by the European Bioinformatics Institute (EBI). The origin research, genome type (isolate or MAG), estimated completeness and contamination, and taxonomic classification are some of the details included in this metadata. These classifications were created before the most recent major taxonomic changes, so it was chosen to reclassify the genomes. Especially since the UHGG collection classification does not separate the Bacteroidaceae and Prevotellaceae families.

To construct the working dataset, all species-level clusters containing *Bacteroides* or *Prevotella* were identified and extracted from the metadata file. From each species cluster, the ten genomes with the highest estimated quality were selected, prioritizing isolates over

```
1. grep -w "$line" $( echo ${path}/data/genomes
   ${family#g__}.tsv) | sort -k2,2 -k7,7nr -k8,8n | head -n 10
   > ${path}/data/tmp.txt
```

MAGs where available. Genome quality was assessed based on the highest completeness over lowest contamination values calculated using CheckM (v1.0.11) (Parks et al., 2015) reported in the metadata (See supplementary material 3.1. UHGG_download.sh). Species clusters with fewer than ten genomes were excluded from the analysis to ensure consistent coverage and allow for reliable comparative analysis (See supplementary material 3.2. filtering.sh)

The same process was used for species clusters that were annotated under the family Bacteroidaceae but were not species-level assigned to either genus, in addition to *Bacteroides* and *Prevotella*. This step was added to take into consideration species that are still taxonomically related to one of the two families of interest but have been reassigned into new genera. Combining these methods yields a dataset that provides a taxonomically inclusive and representative picture of the diversity of Bacteroidaceae and Prevotellaceae in the human gut microbiome.

2.1.2 Taxonomic reclassification with PhyloPhlAn for SGB comparability

As discussed in the previous section, the original taxonomic labels included in the UHGG metadata file were not sufficiently up to date to support a high-resolution comparison of Bacteroidaceae and Prevotellaceae diversity. In particular, the classification scheme used in UHGG version 2.0 does not reflect recent reclassifications at the genus and family levels, and often fails to distinguish between key groups such as *Segatella*, *Prevotella*, and *Bacteroides*. This is due to the usage of an early version of GTDB-Tk (v0.3.1) (Chaumeil et al., 2019). To resolve this, an independent taxonomic reassignment was performed using PhyloPhlAn 3.0 (Asnicar et al., 2020).

PhyloPhlAn is a widely used tool for genome-based phylogenetic analysis and high-resolution microbial taxonomy. It was selected for this step because it allows for standardized, marker gene-based classification of genomes into species-level genome bins (SGBs), including unnamed SGBs (uSGBs) that are not represented in traditional reference databases. Additionally, a key advantage of PhyloPhlAn is that SGB numbers are persistent and stable across database updates, even as genus or species names are revised. This makes analyses based on SGB identifiers robust to future taxonomic changes and ensures compatibility with other tools and resources that also use SGB-based frameworks (Asnicar et al., 2020).

The script phylophlan_assign_sgbs was applied to each of the downloaded genome groups using the updated marker set released in June 2023 (See supplementary material 3.3. assign_sgbs.sh). This marker set incorporates recent taxonomic refinements and supports compatibility with large-scale comparative analyses. The default parameters were used unless otherwise specified, and genomes were classified based on core marker presence and average nucleotide identity thresholds used by the PhyloPhlAn pipeline.

```
1. phylophlan_assign_sgbs /
   -i ${path}/data/Genomes_${family}/FASTA_filtered /
   -o ${path}/results/SGB_1_${family}_${database} /
   -d ${database} /
   --nproc 8 /
   -n 1 /
   --database_folder ${path}/base_files/phylophlan_databases
```

The resulting output included updated SGB assignments for each genome, along with a full taxonomic lineage based on the marker gene phylogeny. These files were subsequently processed using a custom Python script (See supplementary material 3.5. SGB_tables.py), which integrated the following information: the UHGG genome identifier, the original UHGG taxonomic annotation, the newly assigned SGB number, and the updated taxonomic classification. The resulting overview table served as the basis for the diversity analyses in the following sections.

In addition to providing refined taxonomy, this reclassification step ensured consistency across the dataset and enabled comparison of genomes at a standardized species level. By using PhyloPhlAn to generate consistent species-level assignments, this analysis improves both the resolution and the comparability of ecological patterns explored in later chapters.

2.1.3 Phylogenetic validation using the Genome Taxonomy Database (GTDB)

In addition to the reclassification performed with PhyloPhlAn, the Genome Taxonomy Database Toolkit (GTDB-Tk) 2.4.0 (Pierre-Alain Chaumeil et al., 2022) was employed to provide an independent and evolutionarily informed taxonomic framework. While PhyloPhlAn excels in assigning species-level genome bins (SGBs) using a curated set of marker genes, GTDB-Tk uses a genome-wide phylogenetic approach to classify microbial genomes based on metrics such as relative evolutionary divergence (RED) and average nucleotide identity (ANI). This dual approach allowed for both cross-validation of taxonomic assignments and additional insight into evolutionary relationships among the selected genomes.

The GTDB-Tk de novo workflow was applied to all previously selected genome groups (See supplementary material 2.4. GTDB_denovo.sh). The phylum Chloroflexota was identified as an outgroup for the purpose of phylogenetic rooting. Because Chloroflexota is a stable, monophyletic phylum that is sufficiently distinct from Bacteroidota, it was selected. Because of this, it is a suitable point of reference when building rooted phylogenetic trees. A taxonomic filter was used to restrict categorization to genomes within the family Bacteroidaceae, which was the original broad family label before it was recently divided into Bacteroidaceae and Prevotellaceae.

```
1.  gtdbtk de_novo_wf --genome_dir
   ${path}/data/Genomes_${Family}/FASTA_filtered \
                    --taxa_filter f__${Family} \
                    --bacteria \
                    --outgroup_taxon p__Chloroflexota \
                    --out_dir
   ${path}/data/GTDB2_denovo_${Family} \
                    -x .fa \
                    --cpus 20
```

The GTDB-Tk output included full taxonomic lineages for each genome, along with RED values and ANI-based placement within the current GTDB phylogeny. These results were processed using a custom Python script (See supplementary material 3.5. SGB_tables.py), which was designed to integrate multiple sources of taxonomic data. The resulting table included the UHGG genome identifier, the original UHGG classification, the PhyloPhlAn-assigned SGB number, the GTDB-Tk taxonomy, and the updated phylogenetic lineage based on GTDB markers (See supplementary material 2.1. Classification overview). This table was used both for comparative analysis of classification consistency and for visual summaries of species-level diversity across the two target families.

Together, the use of PhyloPhlAn and GTDB-Tk enabled a more comprehensive and robust classification of genomes. By comparing the outputs of both tools, it was possible to identify classification discrepancies, confirm stable taxonomic groups, and ensure that subsequent ecological analyses were based on consistent and evolutionarily coherent species definitions.

2.1.4 Phylogenetic Reconstruction and Visualization

Following the reclassification of genomes using PhyloPhlAn and GTDB-Tk, the newly assigned taxa were placed into a phylogenetic framework to contextualize the observed species-level diversity within each family. This step was essential for validating taxonomic groupings, examining evolutionary relationships, and facilitating subsequent comparative analyses.

PhyloPhlAn offers a quick and scalable method for creating species-level phylogenies by building trees using a carefully chosen set of marker genes. Two distinct maximum-likelihood phylogenies were inferred using the PhyloPhlAn 3.0 workflow: one that included all genomes allocated to the Bacteroidaceae family and another that included all genomes assigned to the Prevotellaceae family. The precision parameter was set to fast for both trees in order to reduce runtime for large datasets, while the diversity parameter was set to medium for genus and family level analyses. (See phylophlan.sh, supplemental material 3.6). This setup was ideal for the project's scope since it struck a balance between phylogenetic resolution and computing efficiency.

```
2. phylophlan -i FASTA_filtered \
       --input_folder
   ${path}/data/Genomes_${family}/FASTA_filtered \
       -o $path/data/${family}_phylophlan \
       --databases_folder
   ${path}/base_files/phylophlan_databases \
       -d phylophlan \
       --diversity medium \
       --fast \
       --configs_folder $path/base_files/ \
       -f supermatrix_aa.cfg \
       --nproc 20 \
       --genome_extension .fa \
       --verbose
```

In parallel, the taxonomic placements generated by the GTDB-Tk v2.4.0 de novo workflow (as described in Section 2.1.3) were used to provide an independent phylogenetic reference. The inclusion of RED values allows GTDB-Tk to place genomes into a rooted, evolutionarily calibrated tree, offering greater resolution and accuracy in the assessment of deep taxonomic relationships. For this reason, GTDB-Tk was considered the primary framework for evolutionary comparisons, while PhyloPhlAn served as the primary source for reclassification.

The marker-based trees from PhyloPhlAn and the RED-informed trees from GTDB-Tk were visualised using the Interactive Tree of Life (iTOL) tool (Letunic & Bork, 2024). These trees were further improved using the List of Prokaryotic Names with Standing in Nomenclature (LPSN) and annotated in line with the updated taxonomic assignments to assure nomenclatural conformity with current classification criteria (Parte et al., 2020). The resulting visualizations can be used to illustrate patterns in taxonomic inconsistencies and clade-level expansion, as well as the overall diversity of each family.

2.2 Investigating Co-occurrence patterns across taxonomic levels

Building on the refined overview of gut microbial diversity established in Section 2.1, this next phase of the analysis focuses on the relationships between microbial taxa. It aims to re-evaluate co-occurrence and co-exclusion patterns in light of recent taxonomic reclassifications and the inclusion of unnamed species-level genome bins (uSGBs), which provide a previously unavailable resolution.

Early research frequently presented *Bacteroides* species and *Prevotella copri* as ecologically and functionally contrasting, as discussed in Chapter 1.3. Lifestyle differences were often linked to this dichotomy: *Bacteroides* species were linked to Westernized dietary patterns high in fat and protein. *P. copri* was typically discovered in those who lived a Non-Westernized lifestyle with fiber-rich diets. The emergence of genome-resolved metagenomics and the incorporation of uSGBs have uncovered a more intricate picture of microbial diversity, as discussed in Chapter 1.5. These developments cast doubt on the simplicity of previous models and raise the possibility that observable community structures are the result of more complex interaction patterns.

Co-occurrence patterns extend beyond species-level dynamics. They can be found at higher taxonomic levels and are often impacted by shared ecological roles, metabolic compatibility, or exclusionary limitations. Moreover, the markup of the gut microbiome can also be studied using dynamic community-level constructs such as microbial networks, co-abundance clusters, or enterotype-like groups.

To investigate these updated co-occurrence patterns, a curated metagenomic dataset was selected that includes high-quality sequencing data and metadata. The inclusion of associated lifestyle factors was essential for ensuring reproducibility and for enabling comparisons between Westernized and Non-Westernized microbiomes.

Community composition was profiled using a computational tool capable of assigning taxonomic identities at the level of uSGBs. This ensured a high resolution. The resulting dataset was cleaned, standardized, and prepared for the downstream analyses, including statistical testing, correlation-based co-occurrence assessment, and microbial network inference.

2.2.1 Extraction of relevant metagenomic samples

To investigate species co-occurrence patterns in relation to lifestyle, a representative and well-structured metagenomic dataset was required. For this purpose, the curatedMetagenomicData package (version 1.7.5) (Pasolli et al., 2017), part of the Bioconductor project, was used. This package provides access to a collection of publicly available metagenomic samples, accompanied by standardized metadata. Its structure enables reproducible filtering across large-scale studies and is widely used for population-level gut microbiome research.

```
1. sampleMetadata |>
     filter(non_westernized == "yes") |>
     filter(body_site == "stool") |>
     filter(age_category == 'adult') |>
     datatable(options = list(dom = "t"), extensions =
   "Responsive")
```

Filtering criteria were applied to include only metagenomic samples derived from stool specimens of healthy adult individuals (See supplementary material 3.7. curatedMetagenomicData.R). This was done to minimize effects related to age, health status, or anatomical sampling site. This process identified 14 studies representing Non-Westernized populations and 60 studies categorized as Westernized (See supplementary material 2.2. Included Metagenomes).

To ensure comparability between the two groups, a balanced dataset of 500 samples per category was selected. Equal group sizes were selected to avoid downstream statistical biases in analyses such as abundance profiling, correlation testing, network modeling, and unsupervised clustering (See supplementary material 1.1 Dataset validation). Unequal sampling can distort multivariate comparisons and exaggerate group-specific features due to differences in statistical power.

2.2.2 Calculation of microbial community composition

After assembling a balanced dataset, the next step was to generate species-level taxonomic profiles for each sample. This was achieved using MetaPhlAn 4 (Blanco-Míguez et al., 2023). This computational tool was specifically designed for profiling microbial communities based on unique clade-specific marker genes. MetaPhlAn 4 was selected for its updated marker set, which includes unnamed species-level genome bins (uSGBs) and is aligned with the species binning system used in PhyloPhlAn 3 (Asnicar et al., 2020). This compatibility ensures that the community composition reflects the most recent advances in taxonomic resolution.

The taxonomic classification was performed by executing the metaphlan command with default parameters and the MetaPhlAn 4.1.0 marker database (See supplementary material 3.8. metaphlan.sh). This tool utilized Bowtie2 (Langmead & Salzberg, 2012) for alignment against a curated set of single-copy marker genes. This allowed an efficient and accurate detection of taxa across a wide range of metagenomic samples.

```
1. metaphlan ${read_ID}_1.fastq,${read_ID}_2.fastq /
   --bowtie2out ${read_ID}.bowtie2.bz2 /
   --nproc 8 /
   --input_type fastq /
   -o W_${read_ID}.txt /
   --bowtie2db /lisc/scratch/mirror/metaphlan/4.1.0
```

The output consisted of individual abundance profiles for each sample, which were subsequently merged into an abundance matrix using the merge_metaphlan_tables.py script. This matrix included relative abundances for all detected taxa across the dataset.

2.2.3 Dataset preparation

The final species-level abundance matrix generated by MetaPhlAn was imported into Python for preprocessing prior to downstream analysis. To reduce noise and improve interpretability, a two-step filtering strategy was applied (See supplementary material 3.9. Metaphlan.ipynb).

First, only species that appeared in at least 10% of all samples were kept in the dataset. This prevalence threshold helps to exclude taxa that are overly rare, potentially misclassified, or highly individual-specific. This filtering step is important to exclude species that are unlikely to reflect robust trends.

Second, for a species to be counted within an individual sample, it had to reach a relative abundance of at least 0.5 percent in that specific sample. This abundance threshold was applied to reduce the influence of extremely low-abundance signals that may result from sequencing artifacts or misclassification. Species that fall below this threshold within a given sample were set to zero, effectively filtering them out on a per-sample basis.

These methods contribute to stabilizing network architectures and statistical models, which are sensitive to noise and sparsity. All further steps, like the co-occurrence correlation, microbial network creation, and unsupervised clustering, were based on the filtered table that was produced.

2.3 Comparative analysis of abundance and diversity in Westernized vs. Non-Westernized populations

To investigate whether patterns of microbial diversity differ between Westernized and Non-Westernized lifestyles, the analysis focused on the two target families Prevotellaceae and Bacteroidaceae. Specifically, the goal was to assess whether species belonging to either family show differences in diversity across the two population groups.

Starting from the filtered species-level abundance matrix, a binary presence–absence matrix was constructed. Each species was marked as present or absent in a sample. This binary simplification was chosen to focus on species richness by testing the number of distinct species observed independent of their relative abundances.

The number of unique species belonging to Prevotellaceae and Bacteroidaceae was counted separately in each sample. This resulted in two richness values per individual. These species counts were used to compare family-level diversity between Westernized and Non-Westernized individuals.

To test whether species richness is significantly different, a one-sided Wilcoxon rank-sum test was performed in R using the wilcox.test() function. The directionality of the test was hypothesis-driven, based on previously described trends: Prevotellaceae richness was expected to be higher in Non-Westernized populations, reflecting long-term exposure to fiber-rich diets, and Bacteroidaceae richness was expected to be greater in Westernized cohorts, associated with high-fat, high-protein diets (See supplementary material 3.10. Boxplots_with_lines.R).

```
1. if (group_name == "Prevotellaceae") {
       wilcox_test_result <- wilcox.test(W_group, NW_group,
   alternative = "less")  # one-sided test, NW > W
     } else if (group_name == "Bacteroidaceae") {
       wilcox_test_result <- wilcox.test(W_group, NW_group,
   alternative = "greater")  # one-sided test, W > NW
     }
```

This comparative approach was used to assess whether previously reported trends in microbial diversity are still holding up when examined using updated taxonomic frameworks and the inclusion of uSGBs. By integrating genome-resolved classifications, this analysis revisits foundational assumptions of lifestyle-driven microbial community structure with higher resolution and broader taxonomic coverage. The results also serve as a baseline for interpreting ecological relationships explored in subsequent co-occurrence and clustering analyses.

2.4 Pairwise Co-occurrence analysis using spearman correlation

To investigate potential ecological interactions and exclusionary dynamics among species in the human gut microbiome, a pairwise correlation analysis was performed on the filtered abundance matrix. This analysis was conducted in R, using the built-in cor() function to compute Spearman rank correlation coefficients between all species pairs.

```
1. cor_matrix <- cor(data_cleaned, method = "spearman")
```

Spearman correlation was selected over Pearson correlation due to its robustness to non-linear relationships and its tolerance for outliers. Both of these features are common in microbiome datasets. Spearman correlation scores the degree and direction of monotonic correlations. This makes it more suitable for identifying co-occurrence and exclusion patterns that do not follow linear trends. Pearson correlation, on the other side assumes normally distributed values and linear relationships. This is especially important when looking at species that might be mutually exclusive, such as those in the Bacteroidaceae and Prevotellaceae families.

The resulting correlation matrix served as a foundational resource for identifying candidate interactions between species, and provided initial evidence for broader ecological structuring patterns within and between these two dominant gut families. This analysis also informed the subsequent construction of microbial co-occurrence networks and clustering approaches.

1.5 Construction of Co-occurrence networks based on species associations

The next step was to acquire an overview of the broader structure of microbial interactions across the entire dataset. The aim of this step was to generate a network-level perspective on species co-occurrence and exclusion.

```
1. net_pears <- netConstruct(cor_matrix,
                             dataType = "correlation" ,
                             sparsMethod = "threshold",
                             thresh = 0.3,
```

The NetCoMi package (Peschel, 2025) in R was used to show the correlation results as a microbial co-occurrence network. A correlation criterion of 0.3 was used for both positive and negative values. This cutoff was chosen to highlight patterns that are meaningful while reducing the inclusion of weak or possibly wrong correlations, which are frequently found in high-dimensional microbiome data.

In addition to offering insight into potential mutualistic or antagonistic relationships, the structure of the network also informed the parameter selection for the following clustering analysis.

1.6 Clustering analysis to identify microbial groupings across samples

The previous steps focused on species-level interactions within the human gut microbiome. The next step shifts the focus from pairwise associations to broader patterns. This clustering analysis is linked to the idea of enterotypes, that is discussed in Chapter 1.7.2. The goal is to reframe the question to include updated taxonomies and the inclusion of previously uncharacterized taxa.

Rather than aiming to reproduce rigid enterotype categories, this approach seeks to re-evaluate dominant community interactions by incorporating the expanded microbial diversity that is available through genome-resolved metagenomics. In doing so, the analysis aims to determine whether formerly overlooked taxa contribute to a clearer or more functionally meaningful understanding of gut microbiome structure.

Key criticisms of the original enterotype model were considered. To account for limited reproducibility and sensitivity to cohort composition, this study used a large, balanced, and lifestyle-diverse dataset. This allowed a more generalizable clustering analysis.

Spectral Clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN), were used as complimentary clustering techniques. This was decided to handle the high-dimensional and non-linear character of microbiome data. DBSCAN was chosen because it can detect clusters of any shape and sort out noise. Spectral clustering was selected due to its ability to identify subtle structure in high-dimensional datasets. Both techniques were used alongside automated operations and manual parameter selection.

To further ensure the robustness of the clustering results, a random permutation approach was implemented. This strategy is a control to test whether the identified clusters could show biological patterns or just random variation in the dataset. By using multiple clustering strategies and validation techniques, this multi-step approach strengthens the reliability of the results.

2.6.1 Data preparation

The dataset was prepared using an identical pipeline for all methods. The preparation steps included standardization of species abundances using the StandardScaler function from the Scikit-learn package (Pedregosa et al., 2011), computation of pairwise correlations using the spearmanr function from the SciPy package (Virtanen et al., 2020), and construction of similarity and distance matrices.

The abundances of each species were standardized to a zero mean and unit variance to prevent species with high abundances from disproportionately influencing the clustering results. Let x denote the filtered species-level abundance.

$$\mathrm{z} = \frac{\mathrm{x} - \mathrm{u}}{\mathrm{s}}$$

where u is the mean abundance of species, and s is the standard deviation (scikit-learn, 2019). This transformation ensures that each species contributes equally to distance-based calculations.

Next, pairwise Spearman rank correlations were computed between all species to capture monotonic relationships in the microbial community:

$$r_S = 1 - \frac{6 \sum d_i^2}{n^3 - n},$$

where $d_i$denotes the difference of each pair (Zar, 2005). Spearman correlation was chosen due to its robustness to non-linear relationships, which are common in microbiome data (Weiss et al., 2016). The resulting correlation matrix $[r_{jk}]$was then transformed into a similarity matrix $[s_{jk}]$by rescaling to the interval $[0,1]$:

$$s_{jk} = \frac{\rho_{jk} + 1}{2}.$$

Finally, a distance matrix $[d_{jk}]$suitable for similarity-based clustering was derived from the similarity matrix:

$$d_{jk} = 1 - s_{jk}.$$

These similarity and distance matrices served as standardized inputs for the DBSCAN and Spectral Clustering algorithms. This ensured that both methods operated on an identical representation of the data.

2.6.2 DBScan

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering algorithm that groups data points by spatial proximity and identifies outliers as noise (Ester et al., 1996). It is well-suited for microbiome data, which is often characterized by irregular cluster shapes, sparsity, and varying densities. DBSCAN does not require pre-specifying the number of clusters and is robust to noise. Therefore, it is an effective choice for uncovering natural groupings within complex datasets.

The implementation used in this study was the DBSCAN function from the Scikit-learn package (Pedregosa et al., 2011). It requires two key parameters: epsilon (ε) as the maximum distance between two points to be considered neighbors, and min_samples as the minimum number of neighboring points required to form a dense region.

To guide the parameter selection, the NearestNeighbors function from Scikit-learn was used to compute the distance from each data point to its k-th nearest neighbor. The parameter k was varied with different candidate values for min_samples. The resulting distances were visualized using k-distance plots that show inflection points. These regions are often referred to as "elbows" indicate natural density thresholds. The inflection points were used to select epsilon values that reflect structural boundaries within the dataset. This data-driven approach allowed a calibration of DBSCAN parameters.

2.6.3 Validation and stability assessment of DBScan Clusters

The HDBSCAN algorithm was applied as an independent benchmark to refine parameter selection and validate the clustering outcome. Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) does not require manual specification of the epsilon parameter. It searches for the optimal cluster structure by analyzing the density landscape of the dataset. This makes it valuable for testing whether the manually tuned DBSCAN parameters are consistent with the density patterns in the data.

The HDBSCAN function, implemented via the Scikit-learn package (Pedregosa et al., 2011), provided an unsupervised, data-driven clustering solution. These were used as a comparison for evaluating the stability and plausibility of DBSCAN-derived clusters.

Beyond this comparison, random permutation testing was used to assess the biological relevance of the clustering structure. The random.permutation function from the NumPy package (Harris et al., 2020) was used to independently shuffled species abundance values within each sample. This step preserved the abundance distribution of each sample but disrupted the covariance structure across species. In doing so, it eliminated any meaningful co-variation while maintaining compositional features in each sample.

The persistence or disappearance of the cluster structure in these randomized datasets is a test of robustness. If similar clustering patterns are seen after the permutation the original groupings may have arisen by chance. If the permuted data failed to show coherent clusters, it would reinforce the validity of the observed structures.

2.6.4 Spectral clustering

Spectral Clustering is a graph-based algorithm that sorts data by analyzing the eigenstructure of a similarity matrix. It is designed to minimize the normalized cut. This means it balances intra-cluster similarity with inter-cluster dissimilarity. This makes Spectral Clustering powerful for detecting complex, non-convex, and overlapping clusters. These structures are often seen in high-dimensional microbiome datasets.

In this analysis the SpectralClustering function from the Scikit-learn package (Pedregosa et al., 2011) was used. This algorithm requires the number of clusters k to be specified beforehand, so a tuning step was performed to identify an appropriate value.

The dataset was subjected to a range of k-means clustering solutions, varying k across a plausible interval. For each clustering result, the silhouette score was calculated. The silhouette score shows the ratio to which each sample is matched to its own cluster while being poorly matched to neighbouring clusters. Scores close to 1 indicate compact, well-separated clusters, whereas scores near zero suggest overlapping or ambiguous assignments.

The resulting silhouette scores were visualized as a function of k, allowing for the identification of local maxima. This data-driven approach provided a reproducible method for selecting the number of clusters used in the Spectral Clustering analysis.

2.7 Functional characterization of clusters via CAZyme profiling

To conclude the analytical workflow and reflect the shift from species-centric to function-oriented microbiome analysis, the final step involved the functional characterization of the microbial clusters. This was done by profiling carbohydrate-active enzymes (CAZymes) in representative genomes of species found in the clusters.

For each of each species that is part of a cluster, genome datasets were retrieved from the Unified Human Gastrointestinal Genome (UHGG) collection (Almeida et al., 2020). This step followed the same extraction and filtering procedure detailed in Section 2.2.1 to ensure consistency in data sourcing. The selected genomes were reclassified using PhyloPhlAn, as described in Section 2.1.2, to align with the updated taxonomic framework used throughout the study.

Genome annotation was carried out using Prokka (version 1.14.6) (Seemann, 2014), which produced standardized gene annotations and the corresponding protein FASTA files required for downstream CAZyme analysis.

```
1. prokka --kingdom Bacteria /
   --outdir ${path}/data/prokka_${family}_out/${line%.fa} /
   --locustag L${line%.fa} /
   --prefix ${line%.fa}
```

The annotated genomes were then processed using run_dbcan (version 4.1.4) (Zheng et al., 2023), part of the dbCAN3 pipeline for CAZyme annotation. To maintain computational efficiency while preserving sensitivity, only the HMMER-based search against the CAZyme Hidden Markov Model (HMM) database was performed. This mode of analysis reliably identifies CAZyme gene families present in each genome.

By mapping CAZyme profiles across taxonomic groups, this step provides functional insights into the metabolic specializations associated with the different clusters. These profiles offer an additional layer of interpretation, potentially linking microbial

```
1. run_dbcan ${path}/data/Genomes_CAZY/FAA/${line}.faa/
    protein/
   --tools hmmer /
   --stp_cpu 32 /
   -c ${path}/data/Genomes_CAZY/GFF/${line}.gff /
   --cgc_substrate /
   --out_dir ${path}/data/Genomes_CAZY/CAZY/${line}.PUL.SUB /
   --dia_cpu 32 /
   --hmm_cpu 32 /
   --tf_cpu 32 /
```

composition to substrate preference, fiber degradation capability, and broader lifestyle associations.

3 Results

The results of this study are presented in three major sections. Each section corresponds to a central topic within the broader aim of re-evaluating the composition and structure of the human gut microbiome in light of recent advances in genome-resolved taxonomy.

The first section addresses the absence of a comprehensive, species-level overview of microbial diversity within the families Bacteroidaceae and Prevotellaceae. The second section revisits earlier interpretations of gut microbiome variation, with a focus on lifestyle-associated differences between Westernized and Non-Westernized populations. The third section extends the analysis from individual taxa to whole-community structure.

3.1 Taxonomic diversity of Prevotellaceae and Bacteroidaceae in the human gut

By utilizing the Unified Human Gastrointestinal Genome (UHGG) collection, this analysis revealed the presence of previously uncharacterized species within both the Prevotellaceae (46.55% of identified SGBs) and Bacteroidaceae (14,28% of identified SGBs) families. In addition, the analysis identified two potential new members of the *Segatella copri* complex: *Prevotella* SGB1638, and *Prevotella* SGB1635.

3.1.1 Updated diversity metrics highlight previously unclassified species

The PhyloPhlAn 3.0 workflow identified a total of 42 species in the human gut belonging to the Bacteroidaceae family, including four uSGBs (9.52%). Additionally, two species groups (4.76%) that had been previously annotated as *Bacteroides* in the UHGG metadata could not be assigned to any known species-level genome bin (SGB).

In contrast, the Prevotellaceae family displayed 58 species, all of which could be assigned to existing SGBs. Among these, 27 (46,55%) species were classified as uSGBs, reflecting a substantial proportion of previously uncharacterized diversity. Importantly, the analysis successfully recovered nine species belonging to the *Segatella copri* complex that are currently known to occur in the human gut. While the full *S. copri* complex comprises additional species, these nine species represent the subset that has been confirmed in human-associated microbiomes**.**

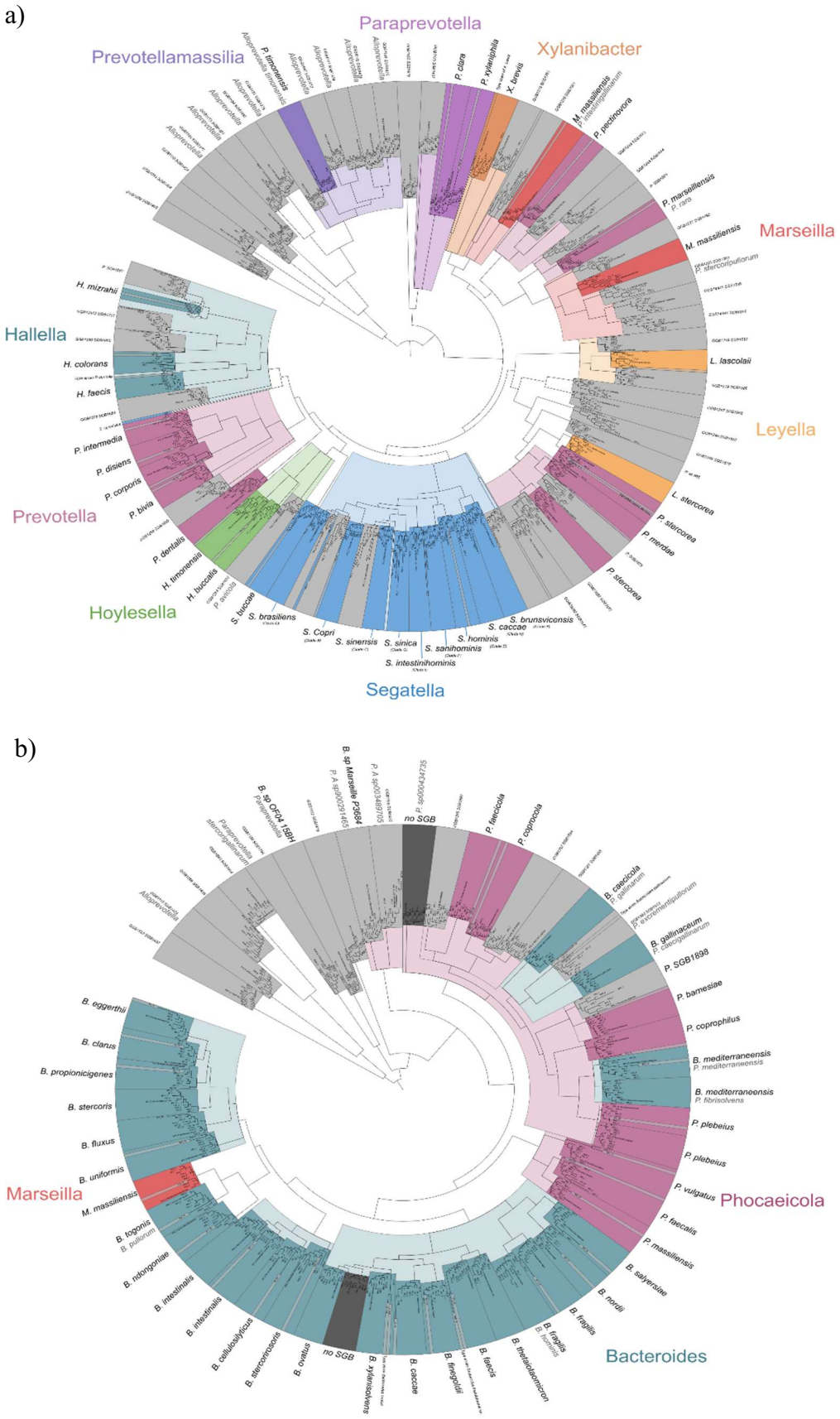


Figure 1 *Phylogenomic reconstruction of Prevotellaceae and Bacteroidaceae species diversity.*

While PhyloPhlAn provides high-resolution taxonomic placement, the tree topologies should not be interpreted as precise evolutionary relationships. These trees are intended to illustrate species representation and relative clustering rather than strict phylogenetic accuracy.
(a) Circular phylogenetic tree of the Prevotellaceae family, constructed using the PhyloPhlAn 3.0 pipeline. Uncharacterized SGBs (uSGBs), shown in grey. Species belonging to the *Segatella copri* complex are highlighted in blue. (b) Circular phylogenetic tree of the Bacteroidaceae family, also generated with PhyloPhlAn 3.0. Two species groups remained unassigned and are shown in black.

The complete overview table including both PhyloPhlAns and GTDBs classifications can be found in the supplementary material (See Supplementary material 2.1. Classification overview Bacteroidaceae and 2.2. Classification overview Prevotellaceae).

3.1.2 GTDB confirms phylogenetic relationships of newly classified families

Using GTDB-Tk the number of genomes assigned to uncharacterized species-level genome bins (uSGBs) increased from 4 (9.52%) to 6 (14.28%) in Bacteroidaceae and from 27 (46.55%) to 37 (63.79%) in Prevotellaceae.

As illustrated in Figure 2, GTDB-Tk situates the analyzed genomes within a comprehensive, pre-calculated bacterial reference tree. Compared to the trees generated with PhyloPhlAn, GTDB trees are more structurally complex, but also more evolutionarily informative. This broader taxonomic viewpoint allows for the identification of deep-branching lineages and a better understanding of the evolutionary relationships within and between families.

Figure 3 shows a more detailed view of the *Segatella copri* complex including the nine members that are known to be found in the human gut and two uSGBs (*Prevotella* SGB1638, and *Prevotella* SGB1635). Aditionally, it shows *Prevotella avicola* being placed between the ScC and *Segatella buccae*.

Together, the complementary outputs from PhyloPhlAn and GTDB-Tk highlight the benefits of integrating multiple classification strategies. While marker-based workflows like PhyloPhlAn offer rapid and precise taxonomic assignments compatible with SGB-based tools, the genome-wide context provided by GTDB-Tk strengthens confidence in lineage-level relationships and helps interpret diversity patterns through an evolutionary lens.

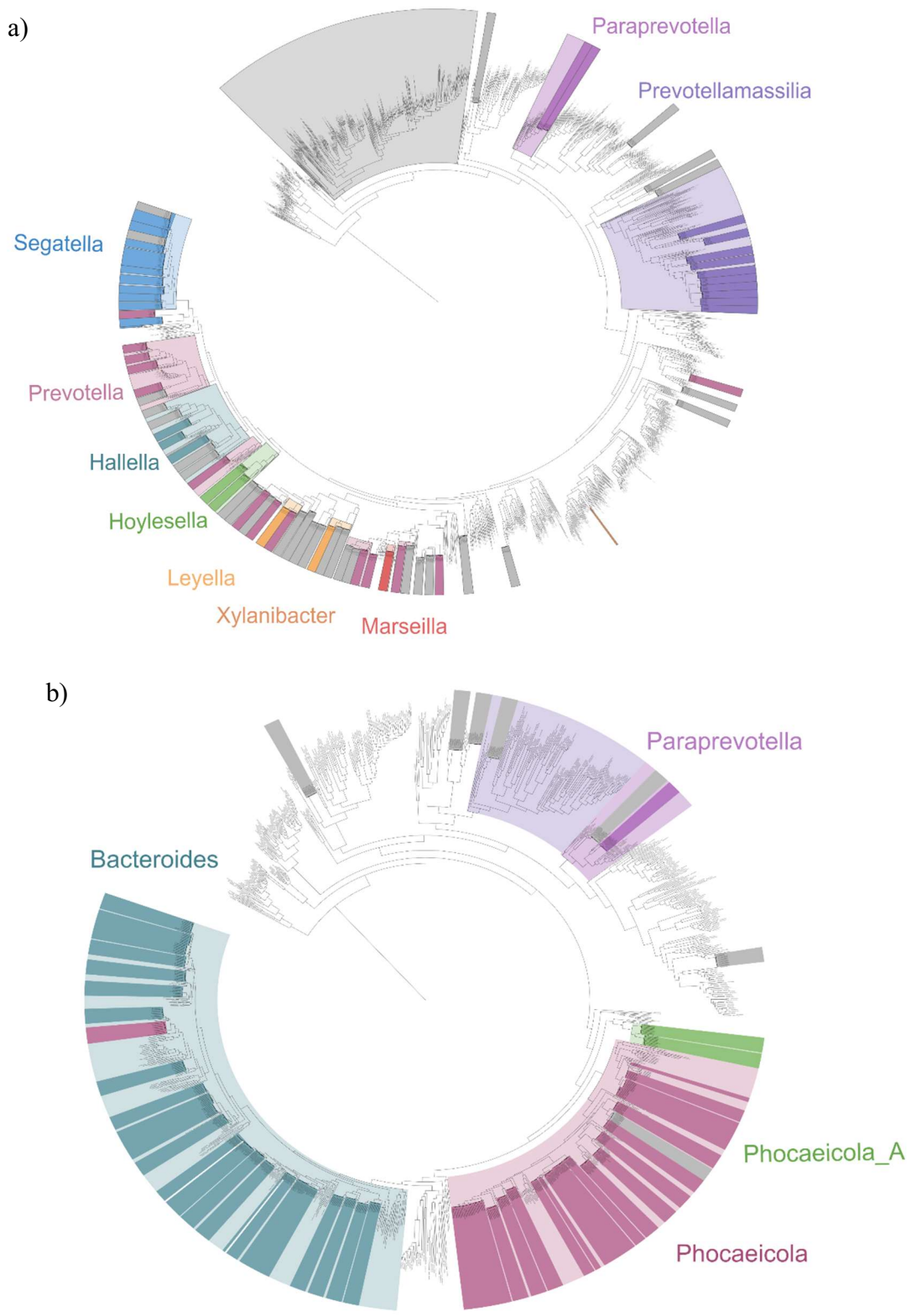


Figure 2 *GTDB-Tk-based phylogenomic placement of Prevotellaceae and Bacteroidaceae genomes.*

(a) Circular reference tree generated using GTDB-Tk, showing the Prevotellaceae family.

(b) Circular reference tree generated using GTDB-Tk, showing the the Bacteroidaceae family,

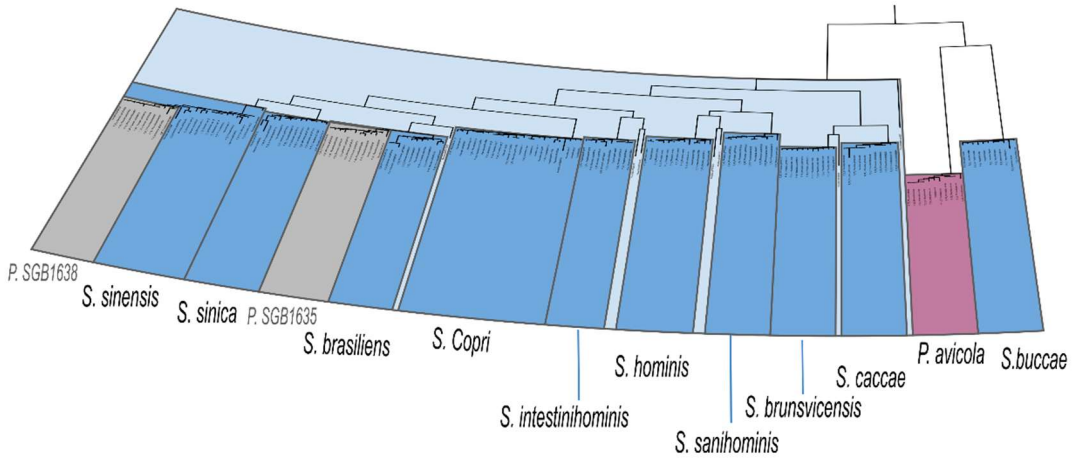


Figure 3 *GTDB-Tk-based phylogenomic placement of the Segatella complex*

Zoomed-in view of the *Segatella copri* clade from Figure 2 panel (a)

3.2 Prevotellaceae species are favored by Non-Westernized lifestyle

Previous research has consistently demonstrated a strong association between lifestyle and gut microbiome composition. It highlighted the dominance of *Prevotella* species in Non-Westernized populations and the prevalence of *Bacteroides* in Westernized contexts. The present study reaffirms these associations using an updated taxonomic framework that includes unnamed species-level genome bins (uSGBs) and expands the analytical scope from individual genera to the broader taxonomic scale of families.

Across all detected species Westernized samples exhibited slightly higher species counts compared to Non-Westernized ones ($p = 0.034$). The taxon-specific analysis showed that Prevotellaceae richness was significantly higher in individuals categorised as Non-Westernized ($p < 2.2 \times 10^{-16}$), while Bacteroidaceae species counts were higher in Westernized samples ($p < 2.2 \times 10^{-16}$).

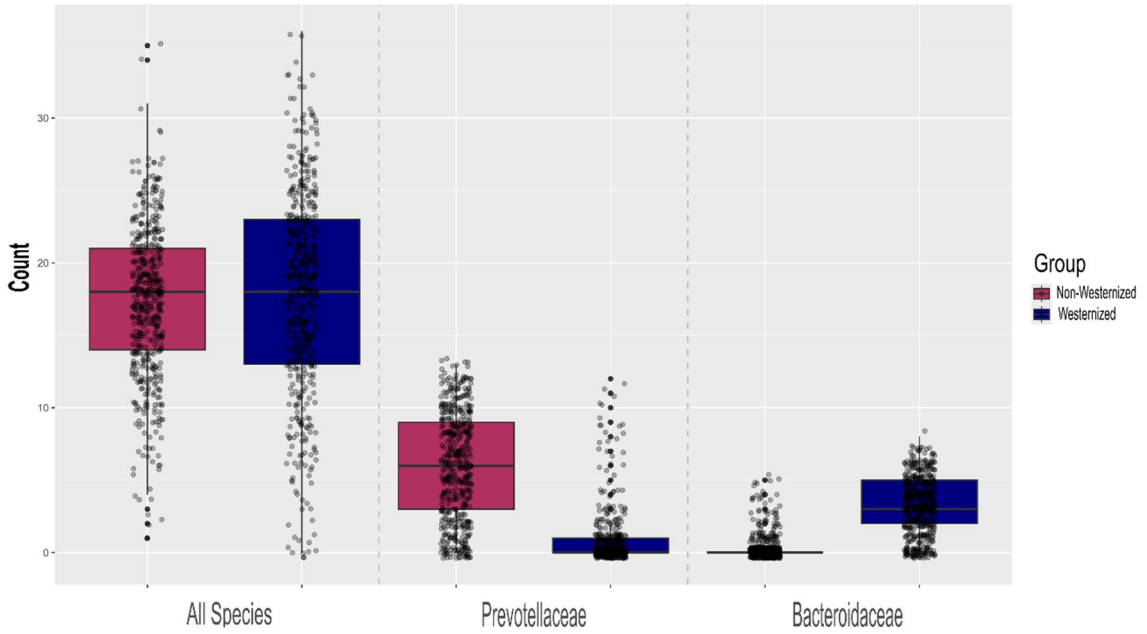


Figure 4 *Lifestyle-associated differences in species richness across total, Prevotellaceae, and Bacteroidaceae taxonomic profiles.*

Boxplots comparing the number of unique species detected per sample in Non-Westernized (red) and Westernized (blue) populations across three categories: all species (left), Prevotellaceae (middle), and Bacteroidaceae (right). Each dot represents a single sample (n = 500 per group). Statistically significant differences were assessed using a non-parametric Wilcoxon rank-sum tests due to violations of normality (Shapiro-Wilk $p < 0.01$ for all comparisons) and unequal variances (F-tests, all $p < 2.2 \times 10^{-16}$).

3.3 Co-occurance patterns of species pairs – Mutual exclusion observed between Prevotellaceae and Bacteroidaceae pairs

Building on the established differences in species diversity between Westernized and Non-Westernized samples, the next step in the analysis focused on understanding how individual taxa co-occur across these lifestyle groups. Historically, *Prevotella copri* has often been described as a key taxon associated with Non-Westernized microbiomes, while *Bacteroides* species are typically linked to Westernized diets. However, more recent classifications, including the division of the former *P. copri* into *Segatella copri* and the broader *Segatella* complex, have shown that no single species can fully explain the observed microbial differences.

This analysis showed that positive correlations are mostly found among species from the same family, particularly within Prevotellaceae, Bacteroidaceae, and Lachnospiraceae. In contrast, the strongest negative correlations occur between species from the genera *Segatella* and either *Phocaeicola* or *Bacteroides*. However, none of the observed negative correlations exceeded -0.5.

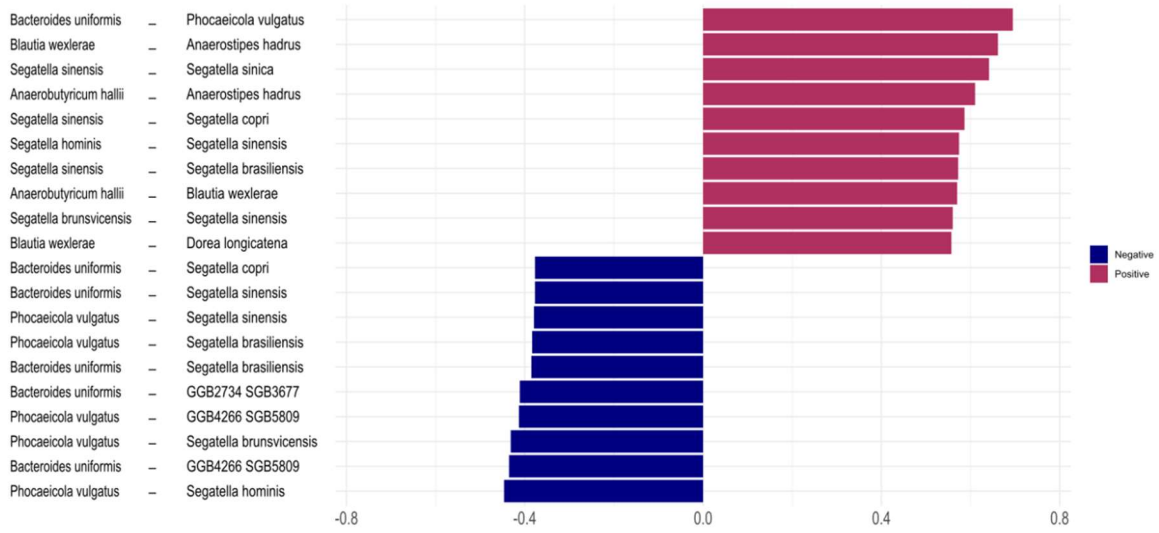


Figure 5 *Pairwise species co-occurrence patterns based on Spearman correlation of relative abundances.*

Barplot showing the top ten strongest positive (red) and negative (blue) pairwise correlations in species abundances across all samples. Species pairs are ranked by the absolute value of their Spearman correlation coefficient, and only statistically significant associations (after multiple testing correction) are included.

This analysis supports a shift in perspective from species-level antagonism to family-level structuring. It also highlights the importance of considering phylogenetic relationships and ecological compatibility when interpreting co-occurrence patterns, especially in the context of lifestyle-associated microbiome variation.

3.4 Tendency of intra-family co-occurrence and inter-family exclusion between Prevotellaceae and Bacteroidaceae

After establishing compositional differences in the human gut microbiome, the next step was to examine species interactions on a broader scale. In this context, network analysis offers a powerful tool for visualizing and interpreting microbial community structure based on co-occurrence and exclusion patterns across the entire dataset.

The resulting species-level co-occurrence network revealed three prominent clusters, each dominated by members of one of the following families: Bacteroidaceae, Prevotellaceae, and Lachnospiraceae. The network highlights dense intra-family connectivity, particularly among Prevotellaceae and Bacteroidaceae members, consistent with the patterns observed in pairwise correlation analyses (see Figure 5). The *Segatella* complex forms a cohesive subnetwork, while negative correlations are frequently observed between members of Prevotellaceae and Bacteroidaceae.

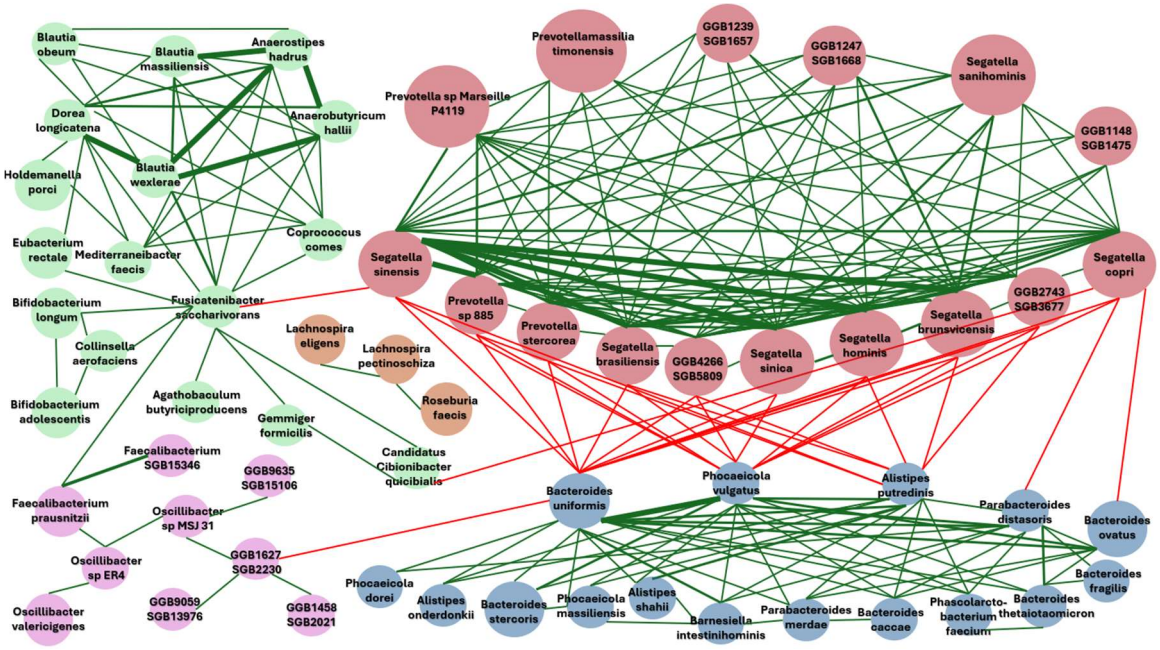


Figure 6 *Network visualization of species-level co-occurrence patterns based on Spearman correlation.*

Microbial co-occurrence network constructed using the NetCoMi package in R, representing statistically significant pairwise correlations between species-level relative abundances. Nodes correspond to individual species, with node colors indicating group affiliation: red (Prevotellaceae), blue (Bacteroidaceae), green (Lachnospiraceae and other Firmicutes). Edge thickness reflects the strength of correlation.

Beyond its ecological implications, the topological features of the network also informed the next stage of the analysis. The family-level clusters guided the parameter selection for the clustering analysis. This helps to ensure that computational approaches to community detection reflected biologically meaningful relationships. In addition to the three major clusters, the network also revealed two smaller but distinct groupings. While these smaller clusters may represent functionally or geographically specialized communities, further investigation would be required to determine their ecological significance.

3.5 Cluster analysis reveals microbial groupings dominated by Prevotellaceae, Bacteroidaceae, and Lachnospiraceae

The preceding network analysis revealed a distinct ecological signal: species co-occurrence patterns appear to be primarily shaped by intra-family interactions, while exclusionary dynamics tend to occur between taxonomic families. This insight provides a more refined framework for understanding microbial community structure and directly informs a re-evaluation of the enterotype concept originally proposed in early gut microbiome research. While the classic enterotype model centered on genus-level groupings by particularly focusing on *Bacteroides*, *Prevotella*, and *Ruminococcus* this study explores whether similar compositional patterns emerge at the family level, incorporating unnamed species-level genome bins (uSGBs) to represent the expanded microbial diversity.

To address the inherent complexity and high dimensionality of microbiome data, two complementary clustering algorithms were employed. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was chosen for its capacity to detect clusters of variable shape and density, and its robustness to outliers which is a critical feature in ecological datasets that often contain sparse or noisy data. In parallel, Spectral Clustering was applied to capture subtler, potentially non-convex structures that may emerge from global similarity patterns, particularly when relationships between taxa are non-linear.

Despite their methodological differences, both clustering approaches converged on a consistent result: clear separation between communities dominated by Prevotellaceae and Bacteroidaceae, along with a third group composed primarily of Lachnospiraceae species. These findings not only echo the structure observed in the network analysis but also reaffirm the ecological distinctiveness of these families under updated taxonomic definitions. Notably, the recovery of such stable clusters across different algorithms and parameterizations adds robustness to the interpretation and supports the hypothesis that family-level community organization is a meaningful and reproducible feature of the human gut microbiome.

3.5.1 DBScan identifies three major clusters reflecting family-level dominance patterns

The clustering results obtained through DBSCAN analysis reveal a clear pattern of family-level dynamics, strengthening insights from both the species co-occurrence network (Figure 6) and the pairwise correlation analysis (Figure 5). Three major clusters were identified, each predominantly composed of species belonging to one of the following families: Bacteroidaceae, Prevotellaceae, and Lachnospiraceae. This consistent grouping not only confirms the validity of the selected DBSCAN parameters (ε = 0.35, MinPts = 4; see supplementary material 1.2. DBSCAN parameters) but also underscores the importance of intra-family structuring in shaping human gut microbial communities.

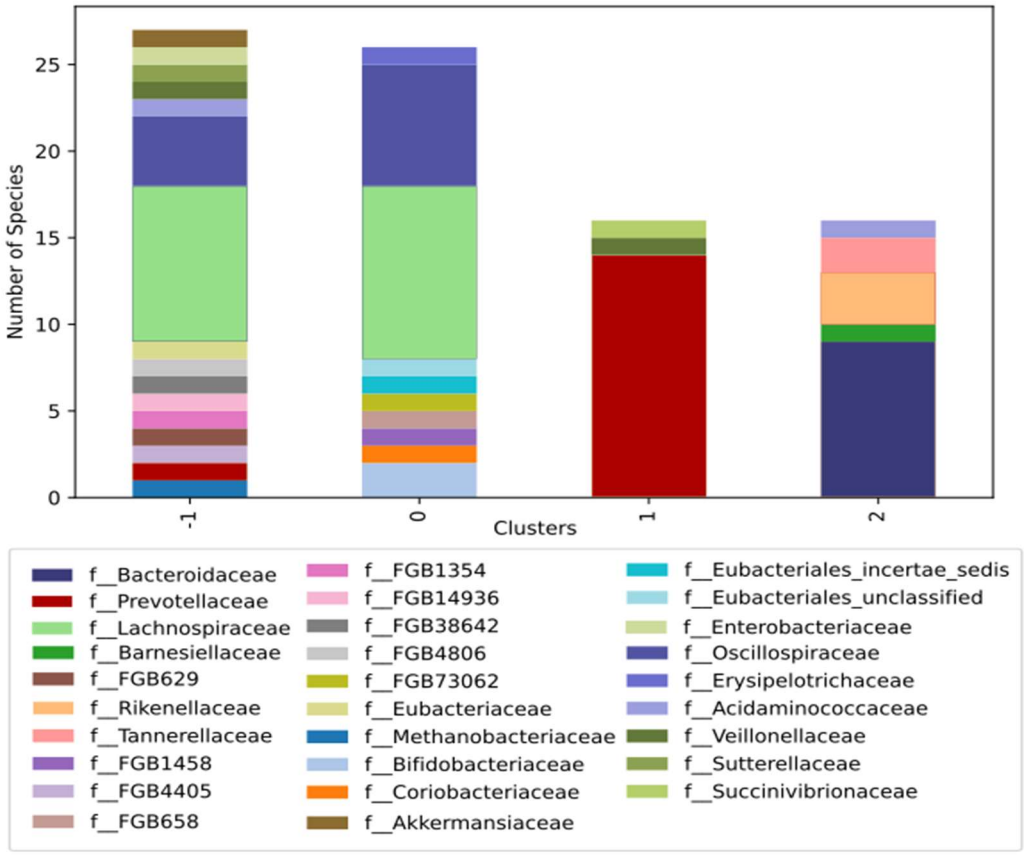


Figure 7 *Family-level composition of DBSCAN-derived clusters.*

Stacked bar plot showing the taxonomic composition of each cluster identified by DBSCAN, based on species-level abundance data. Each bar represents a single cluster (−1 = noise), with colored segments indicating the number of species from different bacterial families. Cluster 1 is composed almost entirely of Prevotellaceae species (dark red), while Cluster 2 is predominantly Bacteroidaceae (dark blue). Cluster 0 is more taxonomically diverse but is heavily enriched in Lachnospiraceae (light green). Species not assigned to any cluster are shown in cluster −1, reflecting outliers or low-density taxa.

3.5.2 HDBScan validates the structure of the three clusters

HDBSCAN is a hierarchical extension of the DBSCAN algorithm that improves its flexibility by estimating the optimal neighbourhood size directly from the data. This approach is particularly advantageous when working with microbial abundance profiles, where density can vary widely between taxonomic groups and across community structures.

Applied to the same dataset, HDBSCAN recovered three main clusters that closely mirrored those identified by DBSCAN. Each cluster was again dominated by species from one of the three major bacterial families: Bacteroidaceae, Prevotellaceae, and Lachnospiraceae. This level of agreement between methods supports the robustness and reproducibility of the clustering structure observed in this study.

One key difference emerged in the treatment of peripheral taxa. HDBSCAN, with its stricter criteria for density-based inclusion, classified a larger number of weakly associated species as noise. In particular, several members of the Oscillospiraceae family were excluded from the Lachnospiraceae cluster and instead marked as unclustered. This adjustment resulted in an even more distinct and taxonomically coherent Lachnospiraceae group, reinforcing the notion that this family forms a tightly interconnected ecological unit.

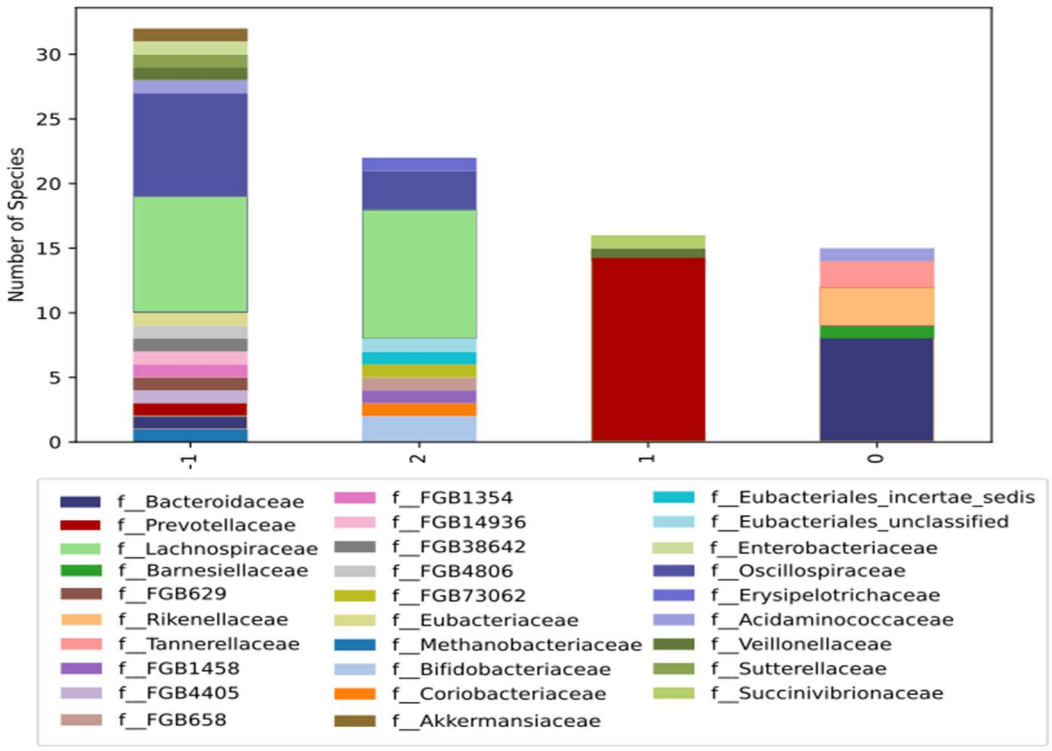


Figure 8 *Family-level composition of HDBSCAN-derived species clusters.*

Stacked bar plot showing the number of species from each bacterial family across clusters identified by HDBSCAN. Each bar represents a cluster, with colors denoting family membership. Cluster 1 is dominated by Prevotellaceae species (dark red), Cluster 0 is primarily composed of Bacteroidaceae (dark blue), and Cluster 2 is enriched in Lachnospiraceae (light green). Species not assigned to any cluster are grouped under cluster −1, with a notable concentration of Oscillospiraceae (blue).

3.5.3 Random permutation confirms family-level dominance within clusters

To further evaluate the stability and robustness of the clustering results, an Adjusted Rand Index (ARI) analysis was performed. The ARI scales the similarity between two clustering solutions. In this context, it was used to compare the original clustering structure to those produced after introducing controlled random perturbations into the dataset. Species-level abundance values were randomly permuted across samples for a defined number of species, while maintaining the original abundance distributions per species.

This procedure tests whether the observed clustering is dependent on a few dominant species or whether it reflects more distributed ecological patterns. When the abundance values of only 5 to 10 species were shuffled, ARI values remained around 0.8. This indicates that the overall clustering structure was preserved, suggesting that the clusters are not overly sensitive to noise or to the influence of individual taxa. As more species were included in the permutation process, the ARI values gradually declined. For example, when 20 to 40 species were shuffled, ARI values dropped to between 0.6 and 0.3, reflecting partial disruption of the original structure. The larger error bars observed in this range suggest that the impact of perturbation varied depending on which species were shuffled, with some contributing more critically to the clustering signal than others.

Once 60 or more species were permuted, ARI values approached zero, indicating a near-complete breakdown of the clustering structure. At this level of perturbation, the ecological signal was lost and the clusters no longer reflected meaningful biological patterns. These findings support the conclusion that the observed clusters are not artifacts of a few dominant species but rather represent broader patterns of co-occurrence and ecological association distributed across multiple taxa. This distributed structure further reinforces the validity of the family-level patterns identified throughout the analysis.

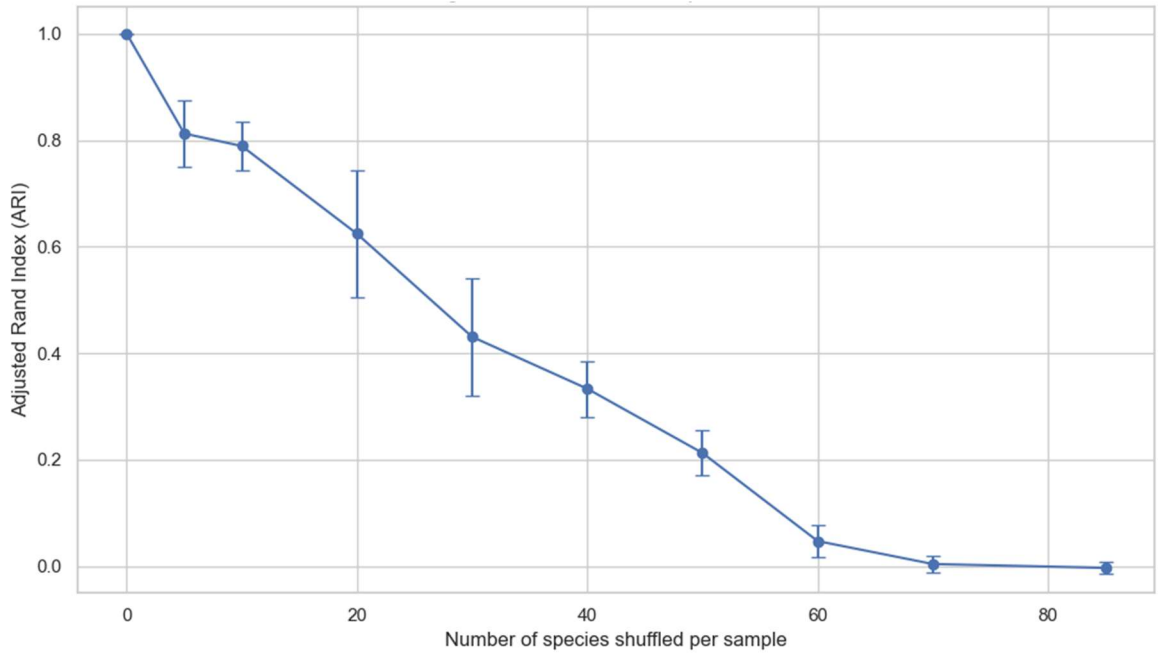


Figure 9 *Robustness of clustering structure assessed through permutation and Adjusted Rand Index (ARI) analysis*

Line plot showing the average Adjusted Rand Index (ARI) values (± standard deviation) resulting from repeated random permutations of species-level abundance data. The x-axis indicates the number of species whose abundance values were shuffled across samples in each permutation run. The y-axis reflects the mean ARI between the original clustering and the clustering obtained after each level of perturbation.

3.5.4 Spectral clustering supports the observed partitioning

Spectral Clustering is particularly well-suited for identifying complex, non-convex, or overlapping clusters. This makes it valuable tool for analysing high-dimensional microbiome data. Spectral Clustering reproduced the main clustering patterns identified in previous analyses, despite lacking an explicit mechanism for detecting outliers. This consistency highlights the algorithm's capacity to uncover underlying structure even when all data points are assigned to clusters.

To determine the optimal number of clusters, silhouette scores were calculated for clustering solutions ranging from 2 to 6 groups. The silhouette coefficient quantifies how well each sample fits within its assigned cluster compared to other clusters, providing an integrated measure of intra-cluster cohesion and inter-cluster separation. These scores were visualized in a silhouette plot (Figure 10), with a red dashed line indicating the average silhouette score for each clustering solution.

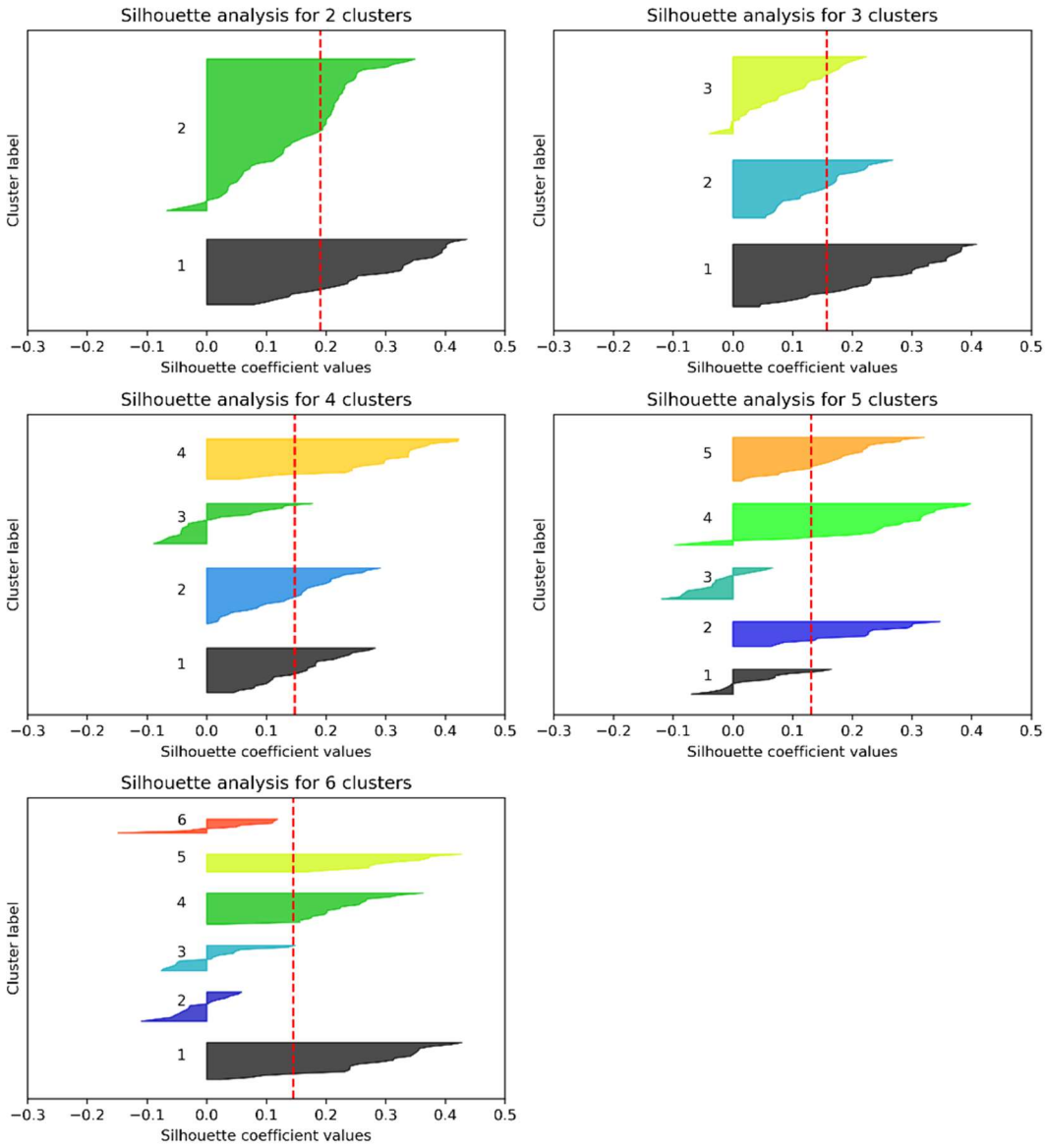


Figure 10 *Silhouette analysis for Spectral Clustering with 2 to 6 cluster solutions.*

Silhouette plots illustrating the distribution of silhouette coefficient values for individual species across different cluster configurations (2 to 6 clusters). Each coloured segment represents the silhouette scores for species assigned to a specific cluster, with the vertical red dashed line indicating the average silhouette score for that clustering iteration.

Overall, the silhouette scores were modest, with the highest score peaking at approximately 0.4. This suggests that while some clustering structure is present, it is not strongly defined. The absence of an outlier-detection step in Spectral Clustering likely contributes to this lower internal consistency, as forcing all species into clusters inevitably leads to the inclusion of weakly associated taxa.

Cluster configurations containing five or six groups were excluded from further consideration due to the emergence of poorly defined clusters, as reflected by below-average silhouette scores and inconsistent internal structure. Among the remaining options, the three-cluster solution was selected as the most robust. This configuration showed the most balanced distribution of cluster sizes and exhibited the least variation in silhouette widths across groups. Notably, this three-cluster structure mirrored the results obtained through DBSCAN and HDBSCAN, offering additional support for the family-level partitioning observed in the gut microbiome.

These results further validate the conclusion that species in this dataset consistently cluster into three distinct ecological groups, each predominantly dominated by a single bacterial family: Prevotellaceae, Bacteroidaceae, and Lachnospiraceae. The convergence of this pattern across multiple clustering methods reinforces the robustness of these community structures and highlights their ecological relevance within the human gut microbiome.

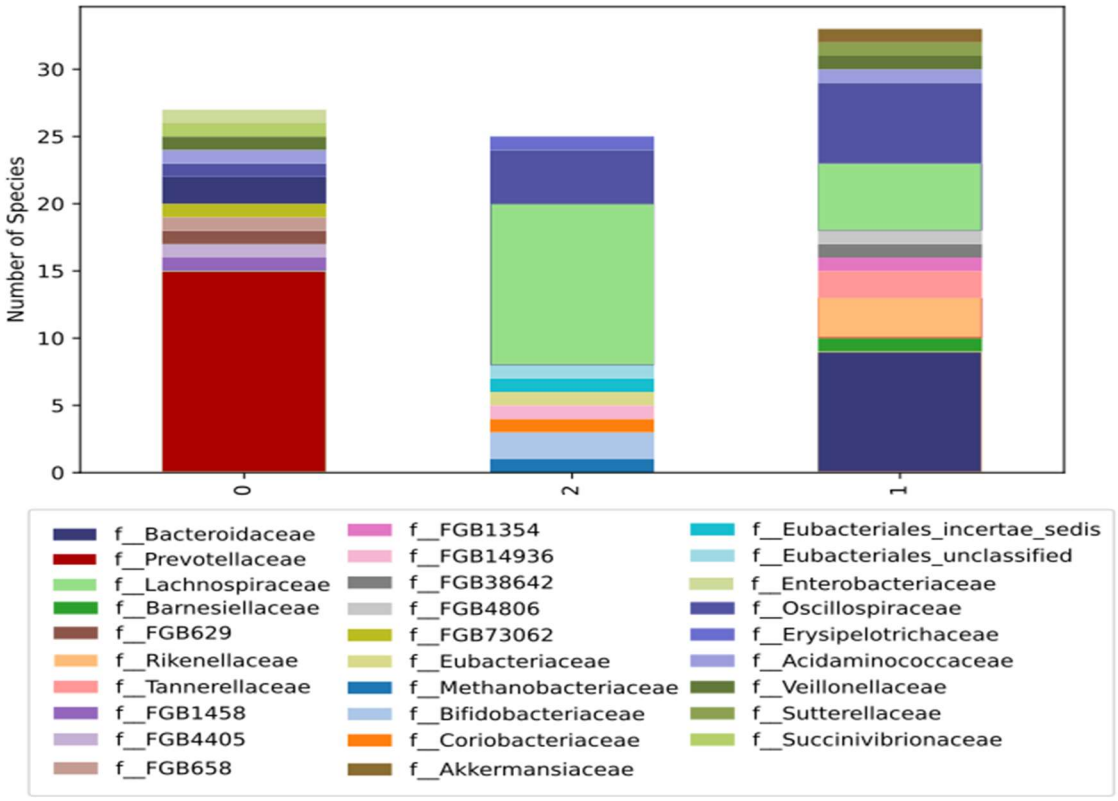


Figure 11 *Family-level composition of species clusters derived from Spectral Clustering.*

Stacked bar plot showing the distribution of species from different bacterial families across three clusters identified via Spectral Clustering. Each bar represents a cluster, and colors indicate family membership. Cluster 1 is dominated by Prevotellaceae species (dark red), Cluster 2 primarily consists of Bacteroidaceae species (dark blue), and Cluster 0 is enriched for Lachnospiraceae species (light green).

3.6 CAZyme Analysis based on Clustering results

The CAZyme profiling revealed functional differentiation among the three major bacterial families dominating the clusters identified through DBSCAN. There was a statistically significant difference in the distribution of abundance-weighted CAZy contribution per species across the three clusters (Kurskal-Wallis p = 0.0034).

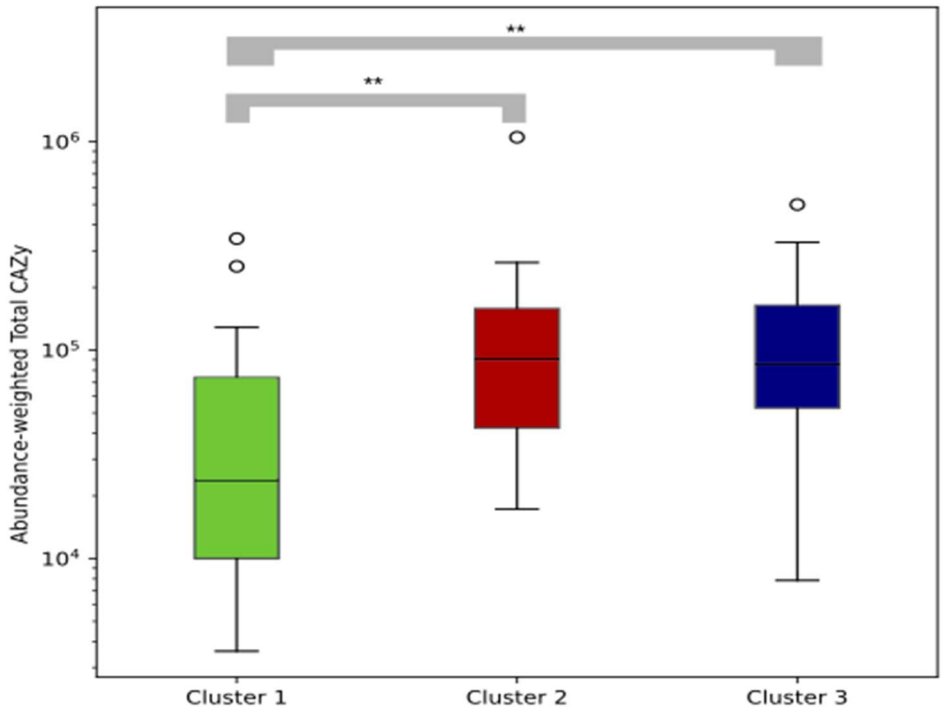


Figure 12 *Abundance-weighted CAZyme distribution across species clusters.*

Boxplots comparing the Abundance-weighted CAZyme distribution across species clusters Cluster 1 is dominated by Lachnospiraceae (green), Cluster 2 by Prevotellaceae (red), and Cluster 3 by Bacteroidaceae (blue).

Cluster 1, dominated by Lachnospiraceae, showed significantly lower weighted CAZyme counts compared to Clusters 1, dominated by Prevotellaceae, and Cluster 2, dominated by Bacteroidaceae (*Mann-Whitney-U* $p < 0.01$). There was no significant difference detected between Cluster 1 and 2.

4 Discussion

4.1 Incorporating SGBs questions the placement of even highly researched groups

Prevotellaceae and Bacteroidaceae are two of the most dominant families in the human gut and a part of the core microbiome (Martínez et al., 2015). Both families are consistently abundant in metagenomic datasets, yet their taxonomic classification has undergone major revisions after the adoption of MAGs and potentially still harbours some taxonomic uncertainty (Pasolli et al., 2019).

One well known example of these revisions is the reorganization of Prevotellaceae. In particular the division from the species *Prevotella copri* into the *Segatella copri* complex, which is part of the new genus Segatella (Tett et al., 2019, Blanco-Míguez, Beghini, et al., 2023). Similar reorganizations can be seen within the Bacteroidaceae family. *Bacteroides dorei* for example got moved into the genus *Phocaeicola* (García-López et al., 2020). These changes show how even dominant and well-characterized families can still be subject to refinement when new technologies enable a deeper look into genetic diversity (Pasolli et al., 2019).

Due to the ongoing taxonomic changes this study focused on tools that use SGB numbers and revised community dynamics on family level. A family-level analysis provides a broader and more stable view since species that were or will be assigned to a new genus are still included. Furthermore, the usage of tools that work with SGB numbers, like MetaPhlAn 4 (Blanco-Míguez et al., 2023) and PhyloPhlAn 3 (Asnicar et al., 2020), provide more stable results. SGB numbers are numerical identifiers that are assigned based on genome similarity thresholds. Therefore, SGB numbers provide a way to identify species even through reassignments and changes of the species name (Pasolli et al., 2019).

The results further underscore that even within well-characterized taxa high-resolution genomic methods can be used to further resolve diversity. Although Bacteroidaceae is widely regarded as one of the most extensively studied bacterial families in the human gut microbiome (Pasolli et al., 2019), the detection of four unnamed species-level genome bins (uSGBs) within this group highlights the capacity of genome-based binning approaches to uncover hidden microbial diversity.

Furthermore, the results suggest that the diversity within the genus *Segatella* could remain incompletely captured and that further investigation will be essential for clarifying the roles and relationships of these candidates. The placement of *Prevotella avicola* between the *Segatella* complex and *Segatella buccae* can also hint at a potential reclassification into the *Segatella* genus.

While Bacteroidaceae appears to be nearing saturation in terms of species discovery within the human gut (9.52% uSGBs), Prevotellaceae remains comparatively under-characterized (46,55%). This is particularly notable given the central role both families play in structuring gut microbiomes across different lifestyles. The finding that a major fraction of Prevotellaceae diversity remains unnamed highlights the importance of continued genome-resolved research.

4.2 Lifestyle-Associated Patterns of Prevotellaceae and Bacteroidaceae

Previous studies showed a dichotomy between Non-Westernized lifestyle favoring *Prevotella* species and Westernized lifestyle *Bacteroides* species (Segata, 2015). This study was able to solidify these findings while incorporating new and unassigned SGBs on a family level. Non-Westernized cohorts showed significantly higher Prevotellaceae richness ($p < 2.2 \times 10^{-16}$), while Bacteroidaceae species counts were higher in Westernized cohorts ($p < 2.2 \times 10^{-16}$). These findings reaffirm that the Prevotellaceae–Bacteroidaceae contrast is a stable and reproducible feature of human gut microbiome variation.

While the common view in literature sees a higher overall diversity being connected to the Non-Westernized lifestyle (Nogal et al., 2021, Manor et al., 2020) this study showed a higher species richness in Westernized samples ($p = 0.034$). While this seems like a contradiction to established knowledge there are several factors that could explain this finding.

A potential difference in sequencing depth was considered during the study design and statistically tested. The sequencing depth between Non-Westernized and Westernized samples was not significantly different ($p = 0.146$) and therefore cannot explain the observed diversity patterns (see Supplementary material 1.1. Dataset validation: sequencing depth is comparable across lifestyle categories despite sampling biases).

Another aspect ties back to the database bias explained in chapter 1.3.3. Microbiome datasets contain fewer samples from Non-Westernized cohorts (Kim et al., 2024, Abdill et al., 2022). This bias could have a strong effect considering the differences in inter- and intra-cohort diversity within these lifestyles as explained by Martínez et al. (2015). Individual Non-Westernized cohorts harbour compositional similar gut communities due to the stable environmental exposure and shared dietary patterns. In contrast, Westernized cohorts tend to show greater variability within individual populations. However, Non-Westernized cohorts from different regions show a higher diversity compared to different Westernized cohorts (Martínez et al., 2015).

In other words, the diversity within cohorts is higher in Westernized samples but diversity between cohorts is higher in Non-Westernized ones (Martínez et al., 2015). To capture these patterns, it is necessary to include more variable Non-Westernized cohorts. It is possible that the chosen dataset failed to account for this dynamic and therefore present a higher overall diversity in Westernized samples.

In conclusion, this study was able to further support the connection between Prevotellaceae and Bacteroidaceae with their contrasting lifestyles. In addition, the results further underlined the impact that the known database bias has on global comparisons.

4.3 Prevotellaceae and Bacteroidaceae species show the expected patterns while Lachnospiraceae species emerge as a new player

The correlation analysis further confirms the decision to consider family-level structures instead of singled out species or groups. This will not only ensure the previously mentioned stability through changes in the naming conventions but also was able to show that dynamics occur within and between certain families (Simonet & McNally, 2021). As shown in previous literature Prevotellaceae species tend to increase together (Tett et al., 2021). Furthermore, it was shown that Prevotellaceae and Bacteroidaceae species tend to co-exclude one another (Yatsunenko et al., 2012). This indicates that the well-known divide between the two families is not driven by any single taxon but reflects a broader split between the families themselves.

Both families also show clear internal structure. Several Prevotellaceae–Prevotellaceae pairs exhibit strong positive correlations (with coefficients above 0.62), and five of the ten strongest correlations in the dataset occur within this family. Bacteroidaceae species show a comparable pattern with the strongest correlation pair being *Bacteroides uniformis* and *Bacteroides vulgatus* ($\rho \approx 0.75$). In addition to these dominant families, four of the ten strongest positive correlations occur between species belonging to the Lachnospiraceae family. These Lachnospiraceae–Lachnospiraceae pairings stand out because they indicate a similar degree of internal coherence within this family.

While these results are able to align with the widely discussed dynamic within and between Prevotellaceae and Bacteroidaceae species, the dynamics of Lachnospiraceae species pairs are still under investigation (Sorbara et al., 2020). A comparative study has pointed out that Lachnospiraceae abundance does not show a connection to Westernized or Non-Westernized lifestyles (Vacca et al., 2020). It is reported that some species of this family are generally enriched in nomadic or hunter-gatherer populations but at the same time typically reduced in the Hadza population (Vacca et al., 2020). Another study has shown that some Lachnospiraceae species tend to be strongly affected by heritability which could be an explanation for the high positive correlation values between Lachnospiraceae pairs (Matsumoto et al., 2021).

In summary, even though it is generally considered that closely related species tend to compete for resources, Bacteroidaceae, Prevotellaceae, and Lachnospiraceae species seem to increase when other members of their own family are present. A co-exclusion is mostly shown between Bacteroidaceae and Prevotellaceae members. This further shows the importance of looking at the gut microbiome at family level.

4.4 Population-Level Gradients further emphasise the Role of Lachnospiraceae

The network analysis was done to further identify patterns within and between families on a broader level. The results show the established dynamic between Prevotellaceae and Bacteroidaceae, which each family forming their own cluster. Furthermore, a third cluster was seen dominated by Lachnospiraceae species. This cluster is not a midpoint between Prevotellaceae and Bacteroidaceae. It shows its own internal structure, with strong positive relationships among several Lachnospiraceae species and only weak links to the other two families. This aligns with the correlation results.

The clustering algorithms support this three-way structure. DBSCAN and Spectral Clustering recovered the same pattern seen in the network, although the boundaries between groups are gradual rather than sharply separated. This matches the idea that these clusters represent broad population-level tendencies instead of discrete enterotypes (Huse et al., 2012). The Lachnospiraceae-rich cluster was the most variable internally, which fits the metabolic diversity known for this family (Sorbara et al., 2020).

Abundance-weighted CAZyme contributions differed significantly between species clusters at the cohort level (Kruskal–Wallis test, $p = 0.003$). Post-hoc pairwise comparisons revealed that the Lachnospiraceae-dominated cluster differed significantly from both the *Prevotella*- and *Bacteroides*-dominated clusters (FDR-adjusted $p < 0.01$), while no significant difference was observed between the *Prevotella*- and *Bacteroides*-dominated clusters.

This suggests a community profile with its own characteristic traits, distinct from both the fiber-adapted *Prevotella*-rich communities typically associated with Non-Westernized lifestyles and the *Bacteroides*-rich communities more common in Westernized populations (Wu et al., 2011). The position of the Lachnospiraceae cluster aligns with previous reports describing members of this family as important butyrate producers that occupy a different functional role from the two dominant families (Vacca et al., 2020).

Additionally, Lachnospiraceae species are often found in connection to Mediterranean lifestyles (Rosés et al., 2021, Latorre-Pérez et al., 2021, Vázquez-Cuesta et al., 2024). This can rais the idea of introducing a third group when looking at different lifestyles especially since the influence of the mediterranean lifestyle has become a new focus in current research (Vázquez-Cuesta et al., 2024).

The CAZyme analysis is supposed to provide a first window into the functional potential of these clusters. Due to the limitations of this work the analysis was kept at a surface level and can easily be influences by a database bias. This is a possible explanation for the lack of significant differences observed between the *Prevotella*- and *Bacteroides*-dominated clusters.

This analysis was done at cohort-level combining samples from individuals whose lifestyle was classified as Westernized and Non-Westernised. Genome databases are biased towards Westernized samples as discussed in chapter 1.3.3 (Kim et al., 2024, Abdill et al., 2022). Therefore, a CAZyme database that is based on a biased genome database will carry on this bias.

4.5 Conclusion

Overall, this thesis demonstrates how genome-resolved approaches refine our understanding of gut microbiome structure, even within families that are traditionally considered well characterized. By focusing on Prevotellaceae and Bacteroidaceae while including SGBs, this work provides an updated overview of their diversity while remaining robust to ongoing taxonomic revisions. The detection of a large fraction of unnamed diversity within Prevotellaceae contrasted with the near saturation observed for Bacteroidaceae. This highlights how unevenly microbial diversity has been captured across core gut taxa. These findings underscore that the adoption of SGB-based frameworks is not a technical refinement, but a necessary step for generating stable insights in a field where nomenclature and classification continue to evolve.

In addressing lifestyle-associated patterns and co-occurrence dynamics, this study confirms the robustness of the Prevotellaceae–Bacteroidaceae contrast across Westernized and Non-Westernized populations. Furthermore, this study extended this framework by revealing the population-level coherence of Lachnospiraceae as a third, functionally distinct cluster. Rather than supporting discrete enterotypes, the results point toward a family-level gradient shaped by lifestyle and functional potential.

At the same time, the observed discrepancy in diversity estimates and functional profiles emphasize the influence of database and sampling biases, particularly for Non-Westernized populations. Taken together, these results suggest the inclusion of genome-informed community structures and highlight the need for more inclusive reference databases and geographically diverse sampling to fully resolve global patterns of the gut microbiome diversity.

5 References

Aakko, J., Pietilä, S., Toivonen, R., Rokka, A., Mokkala, K., Laitinen, K., Elo, L., & Hänninen, A. (2020). A carbohydrate-active enzyme (CAZy) profile links successful metabolic specialization of Prevotella to its abundance in gut microbiota. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-69241-2

Abdill, R. J., Adamowicz, E. M., & Blekhman, R. (2022). Public human microbiome data are dominated by highly developed countries. *PLOS Biology*, *20*(2), e3001536. https://doi.org/10.1371/journal.pbio.3001536

Abdill, R. J., Graham, S. P., Rubinetti, V., Ahmadian, M., Hicks, P., Chetty, A., McDonald, D., Ferretti, P., Gibbons, E., Rossi, M., Krishnan, A., Albert, F. W., Greene, C. S., Davis, S., & Blekhman, R. (2025). Integration of 168,000 samples reveals global patterns of the human gut microbiome. *Cell*. https://doi.org/10.1016/j.cell.2024.12.017

Afolayan, A. O., Ayeni, F. A., Moissl-Eichinger, C., Gregor Gorkiewicz, Halwachs, B., & Christoph Högenauer. (2019). Impact of a Nomadic Pastoral Lifestyle on the Gut Microbiome in the Fulani Living in Nigeria. *Frontiers in Microbiology*, *10*. https://doi.org/10.3389/fmicb.2019.02138

Alanzi, A. R. (2024). Exploring Microbial Dark Matter for the Discovery of Novel Natural Products: Characteristics, Abundance Challenges and Methods. *Journal of Microbiology and Biotechnology*, *35*, 10.4014/jmb.2407.07064. https://doi.org/10.4014/jmb.2407.07064

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. (2020a). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, *39*(1). https://doi.org/10.1038/s41587-020-0603-3

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. (2020b). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, *39*(105–114). https://doi.org/10.1038/s41587-020-0603-3

Arif, S. J., Graham, S. P., Abdill, R. J., & Blekhman, R. (2025). Analyzing human gut microbiome data from global populations: challenges and resources. *Trends in Microbiology*. https://doi.org/10.1016/j.tim.2025.05.008

Armstrong, G. E., Rahman, G., Martino, C., McDonald, D., Gonzalez, A., Gal Mishne, & Knight, R. (2022). Applications and Comparison of Dimensionality Reduction Methods for Microbiome Data. *Front. Bioinform*, *2*. https://doi.org/10.3389/fbinf.2022.821861

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., & Kurokawa, K. (2011). Enterotypes of the human gut microbiome. *Nature*, *473*(7346), 174–180. https://doi.org/10.1038/nature09944

Asnicar, F., Thomas, A. M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., Zhu, Q., Bolzan, M., Cumbo, F., May, U., Sanders, J. G., Zolfo, M., Kopylova, E., Pasolli, E., Knight, R., Mirarab, S., Huttenhower, C., & Segata, N. (2020). Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-16366-7

Bensch, H. M., Tolf, C., Waldenström, J., Lundin, D., & Zöttl, M. (2022). Freeze-drying can replace cold-chains for transport and storage of fecal microbiome samples. *PeerJ*, *10*, e13095–e13095. https://doi.org/10.7717/peerj.13095

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C. C., Charles, T., Chen, X., Cocolin, L., Eversole, K., Corral, G. H., Kazou, M., Kinkel, L., Lange, L., Lima, N., Loy, A., Macklin, J. A., Maguin, E., Mauchline, T., McClure, R., & Mitter, B. (2020). Microbiome definition re-visited: old concepts and new challenges. *Microbiome*, *8*(1). https://doi.org/10.1186/s40168-020-00875-0

Bindels, L. B., Watts, J. E. M., Theis, K. R., Carrion, V. J., Ossowicki, A., Seifert, J., Oh, J., Shao, Y., Hilty, M., Kumar, P., Hildebrand, F., Lovejoy, C., Wigley, P., Yu, K., Zhang, M., Zhang, T., Walter, J., Desai, M. S., Huws, S. A., & Schriml, L. M. (2025). A blueprint for contemporary studies of microbiomes. *Microbiome*, *13*(1). https://doi.org/10.1186/s40168-025-02091-0

Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., Manghi, P., Dubois, L., Huang, K. D., Thomas, A. M., Nickols, W. A., Piccinno, G., Piperni, E., Punčochář, M., Valles-Colomer, M., Tett, A., Giordano, F., Davies, R., Wolf, J., & Berry, S. E. (2023a). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology*, *41*(11), 1–12. https://doi.org/10.1038/s41587-023-01688-w

Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., Manghi, P., Dubois, L., Huang, K. D., Thomas, A. M., Nickols, W. A., Piccinno, G., Piperni, E., Punčochář, M., Valles-Colomer, M., Tett, A., Giordano, F., Davies, R., Wolf, J., & Berry, S. E. (2023b). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology*, *41*(11), 1–12. https://doi.org/10.1038/s41587-023-01688-w

Blanco-Míguez, Eric J.C. Gálvez, Edoardo Pasolli, Francesca De Filippis, Amend, L., Huang, K. D., Paolo Manghi, Till-Robin Lesker, Riedel, T., Cova, L., Punčochář, M., Andrew Maltez Thomas, Mireia Valles-Colomer, Schober, I., Thomas C.A. Hitch, Clavel, T., Berry, S. E., Davies, R., Wolf, J., & Spector, T. D. (2023). Extension of the Segatella copri complex to 13 species with distinct large extrachromosomal elements and associations with host conditions. *Cell Host & Microbe*, *31*(11), 1804-1819.e9. https://doi.org/10.1016/j.chom.2023.09.013

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., Schulz, F., Jarett, J., Rivers, A. R., Eloe-Fadrosh, E. A., Tringe, S. G., Ivanova, N. N., Copeland, A., Clum, A., Becraft, E. D., Malmstrom, R. R., Birren, B., Podar, M., Bork, P., & Weinstock, G. M. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology*, *35*(8), 725–731. https://doi.org/10.1038/nbt.3893

Carter, M. M., Olm, M. R., Merrill, B. D., Dahan, D., Tripathi, S., Spencer, S. P., Yu, F. B., Jain, S., Neff, N., Jha, A. R., Sonnenburg, E. D., & Hryckowian, A. J. (2023). Ultra-deep sequencing of Hadza hunter-gatherers recovers vanishing gut microbes. *Cell*, *186*(14). https://doi.org/10.1016/j.cell.2023.05.046

Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F., & Kjelleberg, S. (2007). Use of 16S rRNA and rpoB Genes as Molecular Markers for Microbial Ecology Studies. *Applied and Environmental Microbiology*, *73*(1), 278–288. https://doi.org/10.1128/AEM.01177-06

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, *36*(6). https://doi.org/10.1093/bioinformatics/btz848

Chen, T., Long, W., Zhang, C., Liu, S., Zhao, L., & Hamaker, B. R. (2017). Fiber-utilizing capacity varies in Prevotella- versus Bacteroides-dominated gut microbiota. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-02995-4

Clarridge, J. E. (2004). Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. *Clinical Microbiology Reviews*, *17*(4), 840–862. https://doi.org/10.1128/cmr.17.4.840-862.2004

Clemente-Suárez, V. J., Beltrán-Velasco, A. I., Redondo-Flórez, L., Martín-Rodríguez, A., & Tornero-Aguilera, J. F. (2023). Global Impacts of Western Diet and Its Effects on Metabolism and Health: a Narrative Review. *Nutrients*, *15*(12), 2749–2749. https://doi.org/10.3390/nu15122749

Costea, P. I., Hildebrand, F., Arumugam, M., Bäckhed, F., Blaser, M. J., Bushman, F. D., de Vos, W. M., Ehrlich, S. Dusko, Fraser, C. M., Hattori, M., Huttenhower, C., Jeffery, I. B., Knights, D., Lewis, J. D., Ley, R. E., Ochman, H., O'Toole, P. W., Quince, C., Relman, D. A., & Shanahan, F. (2017). Enterotypes in the landscape of gut microbial community composition. *Nature Microbiology*, *3*(1), 8–16. https://doi.org/10.1038/s41564-017-0072-8

Coyte, K. Z., Schluter, J., & Foster, K. R. (2015). The ecology of the microbiome: Networks, competition, and stability. *Science*, *350*(6261), 663–666. https://doi.org/10.1126/science.aad2602

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., Fischbach, M. A., Biddinger, S. B., Dutton, R. J., & Turnbaugh, P. J. (2013). Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, *505*(7484), 559–563. https://doi.org/10.1038/nature12820

De Filippis, F., Pasolli, E., Tett, A., Tarallo, S., Naccarati, A., De Angelis, M., Neviani, E., Cocolin, L., Gobbetti, M., Segata, N., & Ercolini, D. (2019). Distinct Genetic and Functional Traits of Human Intestinal Prevotella copri Strains Are Associated with Different Habitual Diets. *Cell Host & Microbe*, *25*(3), 444-453.e3. https://doi.org/10.1016/j.chom.2019.01.004

De Filippis, F., Pellegrini, N., Laghi, L., Gobbetti, M., & Ercolini, D. (2016). Unusual sub-genus associations of faecal Prevotella and Bacteroides with specific dietary patterns. *Microbiome*, *4*(1). https://doi.org/10.1186/s40168-016-0202-1

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., Collini, S., Pieraccini, G., & Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences*, *107*(33), 14691–14696. https://doi.org/10.1073/pnas.1005963107

Dekaboruah, E., Suryavanshi, M. V., Chettri, D., & Verma, A. K. (2020). Human microbiome: an academic update on human body site specific surveillance and its possible role. *Archives of Microbiology*, *202*. https://doi.org/10.1007/s00203-020-01931-x

Deutschmann, I. M., Lima-Mendez, G., Krabberød, A. K., Raes, J., Vallina, S. M., Faust, K., & Logares, R. (2021). Disentangling environmental effects in microbial association networks. *Microbiome*, *9*(1). https://doi.org/10.1186/s40168-021-01141-7

Faust, K., & Raes, J. (2012). Microbial interactions: from networks to models. *Nature Reviews Microbiology*, *10*(8), 538–550. https://doi.org/10.1038/nrmicro2832

Foxx, A. J., Franco Meléndez, K. P., Hariharan, J., Kozik, A. J., Wattenburger, C. J., Godoy-Vitorino, F., & Rivers, A. R. (2021). Advancing Equity and Inclusion in Microbiome Research and Training. *MSystems*, *6*(5). https://doi.org/10.1128/msystems.01151-21

Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G., Segata, N., & Huttenhower, C. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*, *15*(11), 962–968. https://doi.org/10.1038/s41592-018-0176-y

Garcia, S. L., Buck, M., McMahon, K. D., Hans-Peter Grossart, Eiler, A., & Warnecke, F. (2015). *Auxotrophy and intrapopulation complementary in the "interactome" of a cultivated freshwater model community*. *24*(17), 4449–4459. https://doi.org/10.1111/mec.13319

García-López, M., Parte, A. C., Sardà Carbasse, J., Meier-Kolthoff, J. P., Reimer, L. C., & Göker, M. (2020). List of Prokaryotic names with Standing in Nomenclature (LPSN) moves to the DSMZ. *International Journal of Systematic and Evolutionary Microbiology*, *70*(11), 5607–5612. https://doi.org/10.1099/ijsem.0.004332

Garza, D. R., & Dutilh, B. E. (2015). From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. *Cellular and Molecular Life Sciences*, *72*(22), 4287–4308. https://doi.org/10.1007/s00018-015-2004-1

Giovannoni, S. J., Britschgi, T. B., Moyer, C. L., & Field, K. G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature*, *345*(6270), 60–63. https://doi.org/10.1038/345060a0

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews. Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Gorvitovskaia, A., Holmes, S. P., & Huse, S. M. (2016). Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. *Microbiome*, *4*(1). https://doi.org/10.1186/s40168-016-0160-7

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., & Gérard-Marchant, P. (2020). Array Programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hitch, T. C. A., Bisdorf, K., Afrizal, A., Riedel, T., Overmann, J., Strowig, T., & Clavel, T. (2022). A taxonomic note on the genus Prevotella: Description of four novel genera and emended description of the genera Hallella and Xylanibacter. *Systematic and Applied Microbiology*, *45*(6), 126354. https://doi.org/10.1016/j.syapm.2022.126354

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., & Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, *1*(5). https://doi.org/10.1038/nmicrobiol.2016.48

Hugerth, L. W., Larsson, J., Alneberg, J., Lindh, M. V., Legrand, C., Pinhassi, J., & Andersson, A. F. (2015). Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biology*, *16*(1). https://doi.org/10.1186/s13059-015-0834-7

Huse, S. M., Ye, Y., Zhou, Y., & Fodor, A. A. (2012). A Core Human Microbiome as Viewed through 16S rRNA Sequence Clusters. *PLoS ONE*, *7*(6), e34242. https://doi.org/10.1371/journal.pone.0034242

Jeffery, I. B., Claesson, M. J., O'Toole, P. W., & Shanahan, F. (2012). Categorization of the gut microbiota: enterotypes or gradients? *Nature Reviews Microbiology*, *10*(9), 591–592. https://doi.org/10.1038/nrmicro2859

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., Perry, T., Kao, D., Mason, A. L., Madsen, K. L., & Wong, G. K.-S. . (2016). Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. *Frontiers in Microbiology*, *7*. https://doi.org/10.3389/fmicb.2016.00459

Kim, N., Ma, J., Kim, W., Kim, J., Belenky, P., & Lee, I. (2024). Genome-resolved metagenomics: a game changer for microbiome medicine. *Experimental and Molecular Medicine/Experimental and Molecular Medicine*. https://doi.org/10.1038/s12276-024-01262-7

Knights, D., Ward, T. L., McKinlay, C. E., Miller, H., Gonzalez, A., McDonald, D., & Knight, R. (2014). Rethinking "Enterotypes." *Cell Host & Microbe*, *16*(4), 433–437. https://doi.org/10.1016/j.chom.2014.09.013

Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., Huttenhower, C., & Ley, R. E. (2013). A Guide to Enterotypes across the Human Body: Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets. *PLoS Computational Biology*, *9*(1), e1002863. https://doi.org/10.1371/journal.pcbi.1002863

Landry, M. J., & Ward, C. P. (2024). Health Benefits of a Plant-Based Dietary Pattern and Implementation in Healthcare and Clinical Practice. *American Journal of Lifestyle Medicine*, *18*(5). https://doi.org/10.1177/15598276241237766

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Latorre-Pérez, A., Hernández, M., Iglesias, J. R., Morán, J., Pascual, J., Porcar, M., Vilanova, C., & Collado, L. (2021). The Spanish gut microbiome reveals links between microorganisms and Mediterranean diet. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-01002-1

Layeghifard, M., Hwang, D. M., & Guttman, D. S. (2017). Disentangling Interactions in the Microbiome: A Network Perspective. *Trends in Microbiology*, *25*(3), 217–228. https://doi.org/10.1016/j.tim.2016.11.008

Lee, K., Raguideau, S., Sirén, K., Asnicar, F., Cumbo, F., Hildebrand, F., Segata, N., Cha, C.-J., & Quince, C. (2023). Population-level impacts of antibiotic usage on the human gut microbiome. *Nature Communications*, *14*(1), 1191. https://doi.org/10.1038/s41467-023-36633-7

Lemos, L. N., de Carvalho, F. M., Gerber, A., Guimarães, A. P. C., Jonck, C. R., Ciapina, L. P., & de Vasconcelos, A. T. R. (2021). Genome-centric metagenomics reveals insights into the evolution and metabolism of a new free-living group in Rhizobiales. *BMC Microbiology*, *21*(1). https://doi.org/10.1186/s12866-021-02354-4

Letunic, I., & Bork, P. (2024). Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research*, *52*(W1). https://doi.org/10.1093/nar/gkae268

Li, W.-J., Bhagwan Narayan Rekadwad, Jiao, J.-Y., & Salam, N. (2024). *Exploring Microbial Dark Matter and the Status of Bacterial and Archaeal Taxonomy: Challenges and Opportunities in the Future*. 209–229. https://doi.org/10.1007/978-981-99-5720-0_11

Lipoeto, N. I., Geok Lin, K., & Angeles-Agdeppa, I. (2012). Food consumption patterns and nutrition transition in South-East Asia. *Public Health Nutrition*, *16*(9), 1637–1643. https://doi.org/10.1017/s1368980012004569

Liu, S., Moon, C. D., Zheng, N., Huws, S., Zhao, S., & Wang, J. (2022). Opportunities and challenges of using metagenomic data to bring uncultured microbes into cultivation. *Microbiome*, *10*(1). https://doi.org/10.1186/s40168-022-01272-5

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., Andrews, E., Ajami, N. J., Bonham, K. S., Brislawn, C. J., Casero, D., Courtney, H., Gonzalez, A., Graeber, T. G., Hall, A. B., Lake, K., Landers, C. J., Mallick, H., Plichta, D. R., & Prasad, M. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, *569*(7758), 655–662. https://doi.org/10.1038/s41586-019-1237-9

Manor, O., Dai, C. L., Kornilov, S. A., Smith, B., Price, N. D., Lovejoy, J. C., Gibbons, S. M., & Magis, A. T. (2020). Health and disease markers correlate with gut microbiome composition across thousands of people. *Nature Communications*, *11*(1), 5206. https://doi.org/10.1038/s41467-020-18871-1

Marchesi, J. R., & Ravel, J. (2015). The vocabulary of microbiome research: a proposal. *Microbiome*, *3*(1). https://doi.org/10.1186/s40168-015-0094-5

Martínez, I., Stegen, J. C., Maldonado-Gómez, M. X., Eren, A. M., Siba, P. M., Greenhill, A. R., & Walter, J. (2015). The Gut Microbiota of Rural Papua New Guineans: Composition, Diversity Patterns, and Ecological Processes. *Cell Reports*, *11*(4), 527–538. https://doi.org/10.1016/j.celrep.2015.03.049

Matsumoto, N., Park, J., Tomizawa, R., Kawashima, H., Hosomi, K., Mizuguchi, K., Honda, C., Ozaki, R., Iwatani, Y., Watanabe, M., & Kunisawa, J. (2021). Relationship between Nutrient Intake and Human Gut Microbiota in Monozygotic Twins. *Medicina (Kaunas, Lithuania)*, *57*(3), 275. https://doi.org/10.3390/medicina57030275

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., Aksenov, A. A., Behsaz, B., Brennan, C., Chen, Y., DeRight Goldasich, L., Dorrestein, P. C., Dunn, R. R., Fahimipour, A. K., Gaffney, J., Gilbert, J. A., Gogul, G., Green, J. L., Hugenholtz, P., & Humphrey, G. (2018). American Gut: an Open Platform for Citizen Science Microbiome Research. *MSystems*, *3*(3). https://doi.org/10.1128/msystems.00031-18

Moritz, M., Bruno, J. E., Murphy, D. J., Fernández-Giménez, M. E., & Schareika, N. (2025). Transformations in livestock systems: beyond ranching and pastoralism. *Agriculture and Human Values*. https://doi.org/10.1007/s10460-025-10711-6

Naveed, M., Stojanovska, L., Al Dhaheri, A. S., Cheikh Ismail, L., Apostolopoulos, V., & Ali, H. I. (2022). Lifestyle Interventions for Prevention and Management of Diet-Linked Non-Communicable Diseases among Adults in Arab Countries. *Healthcare*, *11*(1), 45. https://doi.org/10.3390/healthcare11010045

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I-Min., Huntemann, M., Palaniappan, K., Ladau, J., Mukherjee, S., Reddy, T. B. K., Nielsen, T., Kirton, E., Faria, J. P., Edirisinghe, J. N., Henry, C. S., & Jungbluth, S. P. (2020). A genomic catalog of Earth's microbiomes. *Nature Biotechnology*, *39*(4), 499–509. https://doi.org/10.1038/s41587-020-0718-6

Nogal, A., Valdes, A. M., & Menni, C. (2021). The role of short-chain fatty acids in the interplay between gut microbiota and diet in cardio-metabolic health. *Gut Microbes*, *13*(1), 1–24. https://doi.org/10.1080/19490976.2021.1897212

Núñez Casal, A. (2024). Race and indigeneity in human microbiome science: microbiomisation and the historiality of otherness. *History and Philosophy of the Life Sciences*, *46*(2). https://doi.org/10.1007/s40656-024-00614-w

Ozdemir, C., Umut Can Kucuksezer, Ismail Ogulur, Pat, Y., Duygu Yazici, Ardicli, S., Mubeccel Akdis, Nadeau, K., & Akdis, C. A. (2024). Lifestyle Changes and Industrialization in the Development of Allergic Diseases. *Current Allergy and Asthma Reports*, *24*(7), 331–345. https://doi.org/10.1007/s11882-024-01149-7

Parajuli, A., Grönroos, M., Siter, N., Puhakka, R., Vari, H. K., Roslund, M. I., Jumpponen, A., Nurminen, N., Laitinen, O. H., Hyöty, H., Rajaniemi, J., & Sinkkonen, A. (2018). Urbanization Reduces Transfer of Diverse Environmental Microbiota Indoors. *Frontiers in Microbiology*, *9*(84). https://doi.org/10.3389/fmicb.2018.00084

Parfrey, L. W., Walters, W. A., & Knight, R. (2011). Microbial Eukaryotes in the Human Microbiome: Ecology, Evolution, and Future Directions. *Frontiers in Microbiology*, *2:153*. https://doi.org/10.3389/fmicb.2011.00153

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. https://doi.org/10.1101/gr.186072.114

Parra-Soto, S., Tannia Valeria Carpio-Arias, Rios-Castillo, I., Patricio Pérez-Armijo, Landaeta-Díaz, L., Murillo, A. G., Araneda-Flores, J., Cavagnari, B. M., Gómez, G., Morales, G., Cordón-Arrivillaga, K., Miranda-Durán, M., Aguilar, A. M., Ortiz, A., Meza-Miranda, E. R., Nava-González, E. J., Jhon Jairo Bejarano-Roncancio, Núñez-Martínez, B., Lima, M., & Jorge. (2025). Dietary Patterns and Sustainable Lifestyles: A Multicenter Study from Latin America and Spain. *Foods*, *14*(12), 2065–2065. https://doi.org/10.3390/foods14122065

Parte, A. C., Sardà Carbasse, J., Meier-Kolthoff, J. P., Reimer, L. C., & Göker, M. (2020). List of Prokaryotic names with Standing in Nomenclature (LPSN) moves to the DSMZ. *International Journal of Systematic and Evolutionary Microbiology*, *70*(11), 5607–5612. https://doi.org/10.1099/ijsem.0.004332

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M. C., Rice, B. L., DuLong, C., Morgan, X. C., Golden, C. D., Quince, C., Huttenhower, C., & Segata, N. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*, *176*(3), 649-662.e20. https://doi.org/10.1016/j.cell.2019.01.001

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., Beghini, F., Malik, F., Ramos, M., Dowd, J. B., Huttenhower, C., Morgan, M., Segata, N., & Waldron, L. (2017). Accessible, curated metagenomic data through ExperimentHub. *Nature Methods*, *14*(11), 1023–1024. https://doi.org/10.1038/nmeth.4468

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Pierre-Alain Chaumeil, Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2022). GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics*, *38*(23), 5315–5316. https://doi.org/10.1093/bioinformatics/btac672

Pingali, P. (2007). Westernization of Asian diets and the transformation of food systems: Implications for research and policy. *Food Policy*, *32*(3), 281–298. https://doi.org/10.1016/j.foodpol.2006.08.001

Porras, A. M., & Brito, I. L. (2019). The internationalization of human microbiome research. *Current Opinion in Microbiology*, *50*, 50–55. https://doi.org/10.1016/j.mib.2019.09.012

Pudlo, N. A., Urs, K., Crawford, R., Pirani, A., Atherly, T., Jimenez, R., Terrapon, N., Henrissat, B., Peterson, D., Ziemer, C., Snitkin, E., & Martens, E. C. (2022). Phenotypic and Genomic Diversification in Complex Carbohydrate-Degrading Human Gut Bacteria. *MSystems*, *7*(1). https://doi.org/10.1128/msystems.00947-21

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., & Xie, Y. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, *464*(7285), 59–65. https://doi.org/10.1038/nature08821

Rampelli, S., Schnorr, Stephanie L., Consolandi, C., Turroni, S., Severgnini, M., Peano, C., Brigidi, P., Crittenden, Alyssa N., Henry, Amanda G., & Candela, M. (2015). Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. *Current Biology*, *25*(13), 1682–1693. https://doi.org/10.1016/j.cub.2015.04.055

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W.-T., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., & Rubin, E. M. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, *499*(7459), 431–437. https://doi.org/10.1038/nature12352

Rosés, C., Cuevas-Sierra, A., Quintana, S., Riezu-Boj, J. I., Martínez, J. A., Milagro, F. I., & Barceló, A. (2021). Gut Microbiota Bacterial Species Associated with Mediterranean Diet-Related Food Groups in a Northern Spanish Population. *Nutrients*, *13*(2), 636. https://doi.org/10.3390/nu13020636

Röttjers, L., & Faust, K. (2018). From hairballs to hypotheses–biological insights from microbial networks. *FEMS Microbiology Reviews*, *42*(6), 761–780. https://doi.org/10.1093/femsre/fuy030

Salam, N., Xian, W.-D., Asem, M. D., Xiao, M., & Li, W.-J. (2020). From ecophysiology to cultivation methodology: filling the knowledge gap between uncultured and cultured microbes. *Marine Life Science & Technology*, *3*(2), 132–147. https://doi.org/10.1007/s42995-020-00064-w

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G., & Malonia, S. K. (2023). Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology*, *12*(7), 997. https://doi.org/10.3390/biology12070997

scikit-learn. (2019). *StandardScaler*. Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Segata, N. (2015). Gut Microbiome: Westernization and the Disappearance of Intestinal Diversity. *Current Biology*, *25*(14), R611–R613. https://doi.org/10.1016/j.cub.2015.05.040

Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, *5*(209). https://doi.org/10.3389/fpls.2014.00209

Simonet, C., & McNally, L. (2021). Kin selection explains the evolution of cooperation in the gut microbiota. *Proceedings of the National Academy of Sciences*, *118*(6). https://doi.org/10.1073/pnas.2016046118

Sonnenburg, E. D., Smits, S. A., Tikhonov, M., Higginbottom, S. K., Wingreen, N. S., & Sonnenburg, J. L. (2016). Diet-induced extinctions in the gut microbiota compound over generations. *Nature*, *529*(7585), 212–215. https://doi.org/10.1038/nature16504

Sonnenburg, E. D., & Sonnenburg, J. L. (2019). The ancestral and industrialized gut microbiota and implications for human health. *Nature Reviews Microbiology*, *17*(6), 383–390. https://doi.org/10.1038/s41579-019-0191-8

Sorbara, M. T., Littmann, E. R., Fontana, E., Moody, T. U., Kohout, C. E., Gjonbalaj, M., Eaton, V., Seok, R., Leiner, I. M., & Pamer, E. G. (2020). Functional and Genomic Variation between Human-Derived Isolates of Lachnospiraceae Reveals Inter- and Intra-Species Diversity. *Cell Host & Microbe*, *28*(1), 134-146.e4. https://doi.org/10.1016/j.chom.2020.05.005

Stefanie Peschel, (2025). NetCoMi: Network Construction and Comparison for Microbiome Data. R package version 1.2.0. https://netcomi.de NetCoMi

Tannock, G. W. (2023). Understanding the gut microbiota by considering human evolution: a story of fire, cereals, cooking, molecular ingenuity, and functional cooperation. *Microbiology and Molecular Biology Reviews*. https://doi.org/10.1128/mmbr.00127-22

Tett, A., Huang, K. D., Asnicar, F., Fehlner-Peach, H., Pasolli, E., Karcher, N., Armanini, F., Manghi, P., Bonham, K., Zolfo, M., De Filippis, F., Magnabosco, C., Bonneau, R., Lusingu, J., Amuasi, J., Reinhard, K., Rattei, T., Boulund, F., Engstrand, L., & Zink, A. (2019). The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. *Cell Host & Microbe*, *26*(5). https://doi.org/10.1016/j.chom.2019.08.018

Tett, A., Pasolli, E., Masetti, G., Ercolini, D., & Segata, N. (2021). Prevotella diversity, niches and interactions with the human host. *Nature Reviews Microbiology*, *19*(9), 585–599. https://doi.org/10.1038/s41579-021-00559-y

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The Human Microbiome Project. *Nature*, *449*(7164), 804–810. https://doi.org/10.1038/nature06244

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., & Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37–43. https://doi.org/10.1038/nature02340

Vacca, M., Celano, G., Calabrese, F. M., Portincasa, P., Gobbetti, M., & De Angelis, M. (2020). The Controversial Role of Human Gut Lachnospiraceae. *Microorganisms*, *8*(4). https://doi.org/10.3390/microorganisms8040573

Vázquez-Cuesta, S., Lozano García, N., Rodríguez-Fernández, S., Fernández-Avila, A. I., Bermejo, J., Fernández-Avilés, F., Muñoz, P., Bouza, E., & Reigadas, E. (2024). Impact of the Mediterranean Diet on the Gut Microbiome of a Well-Defined Cohort of Healthy Individuals. *Nutrients*, *16*(6), 793. https://doi.org/10.3390/nu16060793

Vinogradova, E., Mukhanbetzhanov, N., Nurgaziyev, M., Jarmukhanov, Z., Aipova, R., Sailybayeva, A., Bekbossynova, M., Kozhakhmetov, S., & Kushugulova, A. (2024). Impact of urbanization on gut microbiome mosaics across geographic and dietary contexts. *MSystems*, *9*(10). https://doi.org/10.1128/msystems.00585-24

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., & Carey, C. J. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://www.nature.com/articles/s41592-019-0686-2

Wang, Z., Usyk, M., Vázquez-Baeza, Y., Chen, G.-C., Isasi, C. R., Williams-Nguyen, J. S., Hua, S., McDonald, D., Thyagarajan, B., Daviglus, M. L., Cai, J., North, K. E., Wang, T., Knight, R., Burk, R. D., Kaplan, R. C., & Qi, Q. (2021). Microbial co-occurrence complicates associations of gut microbiome with US immigration, dietary intake and obesity. *Genome Biology*, *22*(1). https://doi.org/10.1186/s13059-021-02559-w

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L. C., Xu, Z. Z., Ursell, L., Alm, E. J., Birmingham, A., Cram, J. A., Fuhrman, J. A., Raes, J., Sun, F., Zhou, J., & Knight, R. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME Journal*, *10*(7), 1669–1681. https://doi.org/10.1038/ismej.2015.235

Wheaton, F. V., & Crimmins, E. M. (2013). In Hindsight: Urban Exposure Explains the Association Between Prior Migration and Current Health of Older Adults in Mexico. *Journal of Aging and Health*, *25*(3), 422–438. https://doi.org/10.1177/0898264312472537

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y. ., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F. D., & Lewis, J. D. (2011). Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science*, *334*(6052), 105–108. https://doi.org/10.1126/science.1208344

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. A., Lauber, C., Clemente, J. C., Knights, D., & Knight, R. (2012). Human gut microbiome viewed across age and geography. *Nature*, *486*(7402), 222–227. https://doi.org/10.1038/nature11053

Zar, J. H. (2005). Spearman Rank Correlation. *Encyclopedia of Biostatistics*. https://doi.org/10.1002/0470011815.b2a15150

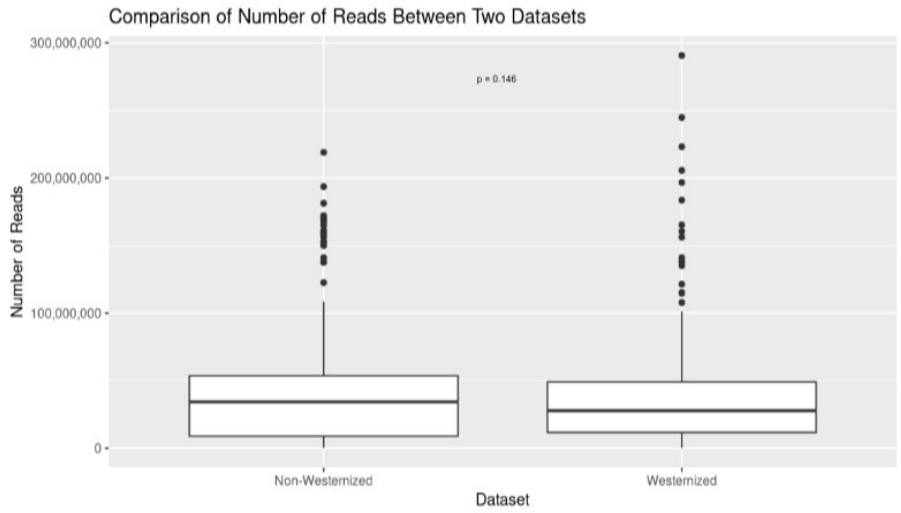
Zheng, J., Ge, Q., Yan, Y., Zhang, X., Huang, L., & Yin, Y. (2023). dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Research*, *51*(W1). https://doi.org/10.1093/nar/gkad328

6 Supplementary Material

Contents

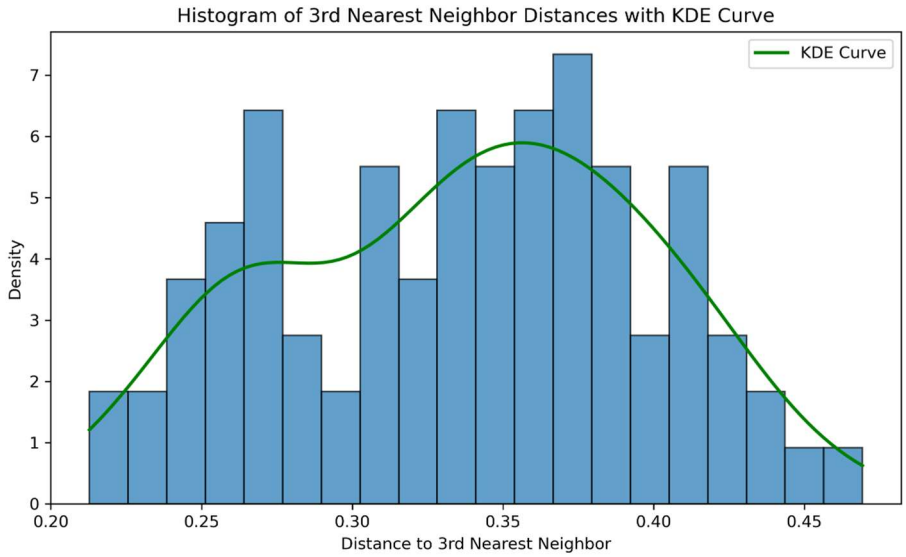
1. Additional Figures

1.1 Dataset validation: sequencing depth is comparable across lifestyle categories *despite sampling biases*



Additional figure 1*: Comparison of sequencing depth between non-Westernized and Westernized metagenomic samples*. Boxplot showing the distribution of total read counts across 500 stool-derived samples per lifestyle category. The Welch's two-sample t-test indicated that this difference was not statistically significant (t = 1.4548, df = 993.64, p = 0.146).

1.2 DBSCAN *parameters*



Additional Figure 2*: Distribution of distances to the 3rd nearest neighbor used for DBSCAN Epsilon calibration*. Histogram of 3rd nearest neighbor distances (corresponding to MinPts = 4) across the dataset, with an overlaid Kernel Density Estimate (KDE) curve.

2. Tabels

2.1 Classification overview Bacteroidaceae

Group_name	UHGG_rep	SGB_ID	%	#down	#avail	%GTDB	Type Strain
Bacteroides_1	MGYG000001780	noSGB	80	10	18	80	
Bacteroides_4	MGYG000002549	SGB1877	100	10	559	100	added
Bacteroides_5	MGYG000002455	SGB1844	100	10	463	100	added
Bacteroides_6	MGYG000000105	SGB1832	100	10	136	100	added
Bacteroides_9	MGYG000001313	SGB1829	100	10	433	100	added
Bacteroides_10	MGYG000002281	SGB1860	100	10	162	100	added
Bacteroides_11	MGYG000000029	SGB1862	100	10	105	100	added
Bacteroides_12	MGYG000001370	SGB1835	100	10	10	100	included
Bacteroides_13	MGYG000001337	SGB1855	100	10	1124	100	added
Bacteroides_14	MGYG000000236	SGB1853	100	10	138	100	not available
Bacteroides_17	MGYG000002470	SGB1846	90	10	156	90	added
Bacteroides_18	MGYG000004019	SGB1824	100	10	37	100	added
Bacteroides_19	MGYG000002621	SGB1821	100	10	23	100	added
Bacteroides_21	MGYG000000265	SGB1858	100	10	66	100	added
Bacteroides_23	MGYG000001378	SGB1871	70	10	1072	70	added
Bacteroides_25	MGYG000001433	SGB1857	100	10	57	100	added
Bacteroides_27	MGYG000000224	SGB1831	100	10	10	100	added
Bacteroides_33	MGYG000000211	SGB1845	90	10	22	90	not available
Bacteroides_38	MGYG000000013	noSGB	100	10	37	100	not available
Bacteroides_40	MGYG000002273	SGB1848	100	10	12	100	added
Bacteroides_41	MGYG000003681	SGB1830	100	10	1843	100	included
Bacteroides_42	MGYG000000196	SGB1861	100	10	579	100	added
Bacteroides_43	MGYG000000652	SGB1822	100	10	57	100	added
Bacteroides_44	MGYG000001346	SGB1836	100	10	6001	100	added
Bacteroides_45	MGYG000001345	SGB1867	100	10	408	100	added
Bacteroides_47	MGYG000003094	SGB1897	100	10	40	100	added
Bacteroides_48	MGYG000001306	SGB1891	100	10	593	100	added
Bacteroides_49	MGYG000000273	SGB1888	100	10	342	100	added
Bacteroides_50	MGYG000002478	SGB1814	80	10	5767	80	added
Bacteroides_52	MGYG000000243	SGB1812	100	10	611	100	added
Bacteroides_53	MGYG000004479	SGB1912	100	9	9	100	added
Bacteroides_54	MGYG000001364	SGB1899	70	10	158	70	added
Bacteroides_55	MGYG000003693	SGB1903	70	10	977	70	not available
Bacteroides_58	MGYG000001835	SGB1895	100	10	111	100	not available
Bacteroides_59	MGYG000001963	noSGB	90	10	119	90	not available
Bacteroides_60	MGYG000003097	SGB1894	100	10	15	100	not available
Bacteroides_61	MGYG000001789	SGB1917	100	10	25	100	added
Bacteroides_63	MGYG000002171	SGB1810	100	10	10	100	added
Bacteroides_65	MGYG000000043	SGB1910	100	10	55	100	not available
Bacteroides_66	MGYG000002933	SGB1914	100	10	18	100	not available
Bacteroides_67	MGYG000000781	SGB1887	100	10	69	100	added
Bacteroides_68	MGYG000002568	SGB1890	100	10	33	100	not available
Bacteroides_71	MGYG000002622	SGB1913	100	9	9	100	not available
Bacteroides_76	MGYG000001787	SGB1898	100	9	9	100	not available

Group_name	Species_UHGG	Species_Phylo	Species_GTDB
Bacteroides_1		no SGB	
Bacteroides_4	Bacteroides caccae	Bacteroides caccae	Bacteroides caccae
Bacteroides_5	Bacteroides cellulosilyticus	Bacteroides cellulosilyticus	Bacteroides cellulosilyticus
Bacteroides_6	Bacteroides clarus	Bacteroides clarus	Bacteroides clarus
Bacteroides_9	Bacteroides eggerthii	Bacteroides eggerthii	Bacteroides eggerthii
Bacteroides_10	Bacteroides faecis	Bacteroides faecis	Bacteroides faecis
Bacteroides_11	Bacteroides finegoldii	Bacteroides finegoldii	Bacteroides finegoldii
Bacteroides_12	Bacteroides fluxus	Bacteroides fluxus	Bacteroides fluxus
Bacteroides_13	Bacteroides fragilis	Bacteroides fragilis	Bacteroides fragilis
Bacteroides_14	Bacteroides fragilis_A	Bacteroides fragilis	Bacteroides hominis
Bacteroides_17	Bacteroides intestinalis	Bacteroides intestinalis	Bacteroides intestinalis
Bacteroides_18	Bacteroides massiliensis_A	Mediterranea massiliensis	Bacteroides massiliensis
Bacteroides_19	Bacteroides ndongoniae	Bacteroides ndongoniae	Bacteroides ndongoniae
Bacteroides_21	Bacteroides nordii	Bacteroides nordii	Bacteroides nordii
Bacteroides_23	Bacteroides ovatus	Bacteroides ovatus	Bacteroides ovatus
Bacteroides_25	Bacteroides salyersiae	Bacteroides salyersiae	Bacteroides salyersiae
Bacteroides_27	Bacteroides sp003545565	Bacteroides propionicigenes	Bacteroides propionicigenes
Bacteroides_33	Bacteroides sp900556215	Bacteroides intestinalis	Bacteroides intestinalis_A
Bacteroides_38	Bacteroides sp902362375	no SGB	Bacteroides sp902362375
Bacteroides_40	Bacteroides stercorirosoris	Bacteroides stercorirosoris	Bacteroides stercorirosoris
Bacteroides_41	Bacteroides stercoris	Bacteroides stercoris	Bacteroides stercoris
Bacteroides_42	Bacteroides thetaiotaomicron	Bacteroides thetaiotaomicron	Bacteroides thetaiotaomicron
Bacteroides_43	Bacteroides togonis	Bacteroides togonis	Bacteroides pullorum
Bacteroides_44	Bacteroides uniformis	Bacteroides uniformis	Bacteroides uniformis
Bacteroides_45	Bacteroides xylanisolvens	Bacteroides xylanisolvens	Bacteroides xylanisolvens
Bacteroides_47	Phocaeicola barnesiae	Phocaeicola barnesiae	Phocaeicola barnesiae
Bacteroides_48	Phocaeicola coprocola	Phocaeicola coprocola	Phocaeicola coprocola
Bacteroides_49	Phocaeicola coprophilus	Phocaeicola coprophilus	Phocaeicola coprophilus
Bacteroides_50	Phocaeicola dorei	Phocaeicola vulgatus	Phocaeicola vulgatus
Bacteroides_52	Phocaeicola massiliensis	Phocaeicola massiliensis	Phocaeicola massiliensis
Bacteroides_53	Phocaeicola mediterraneensis	Bacteroides mediterraneensis	Phocaeicola mediterraneensis
Bacteroides_54	Phocaeicola plebeius	Phocaeicola plebeius	Phocaeicola plebeius
Bacteroides_55	Phocaeicola plebeius_A	Phocaeicola plebeius	Phocaeicola plebeius_A
Bacteroides_58	Phocaeicola sp000432735	GGB1387 SGB1895	Phocaeicola sp000432735
Bacteroides_59	Phocaeicola sp000434735	no SGB	Phocaeicola sp000434735
Bacteroides_60	Phocaeicola sp000436795	GGB1387 SGB1894	Phocaeicola sp000436795
Bacteroides_61	Phocaeicola sp002161565	Bacteroides gallinaceum	Phocaeicola caecigallinarum
Bacteroides_63	Phocaeicola sp002493165	Phocaeicola faecalis	Phocaeicola faecalis
Bacteroides_65	Phocaeicola sp900066455	Bacteroides mediterraneensis	Phocaeicola fibrisolvens
Bacteroides_66	Phocaeicola sp900540105	Bacteroides caecicola	Phocaeicola gallinarum
Bacteroides_67	Phocaeicola sp900541515	Phocaeicola faecicola	Phocaeicola faecicola
Bacteroides_68	Phocaeicola sp900542985	GGB1385 SGB1890	Phocaeicola sp900542985
Bacteroides_71	Phocaeicola sp900546095	Bacteroides sp ET225	Phocaeicola excrementipullorum
Bacteroides_76	Phocaeicola sp900551645	Phocaeicola SGB1898	Phocaeicola sp900551645

Group_name	Lin_UHGG
Bacteroides_1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__
Bacteroides_4	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides caccae
Bacteroides_5	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides cellulosilyticus
Bacteroides_6	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides clarus
Bacteroides_9	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides eggerthii
Bacteroides_10	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides faecis
Bacteroides_11	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides finegoldii
Bacteroides_12	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides fluxus
Bacteroides_13	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides fragilis
Bacteroides_14	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides fragilis_A
Bacteroides_17	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides intestinalis
Bacteroides_18	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides massiliensis_A
Bacteroides_19	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides ndongoniae
Bacteroides_21	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides nordii
Bacteroides_23	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides ovatus
Bacteroides_25	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides salyersiae
Bacteroides_27	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides sp003545565
Bacteroides_33	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides sp900556215
Bacteroides_38	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides sp902362375
Bacteroides_40	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides stercorirosoris
Bacteroides_41	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides stercoris
Bacteroides_42	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides thetaiotaomicron
Bacteroides_43	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides togonis
Bacteroides_44	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides uniformis
Bacteroides_45	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides xylanisolvens
Bacteroides_47	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola barnesiae
Bacteroides_48	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola coprocola
Bacteroides_49	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola coprophilus
Bacteroides_50	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola dorei
Bacteroides_52	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola massiliensis
Bacteroides_53	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola mediterraneensis
Bacteroides_54	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola plebeius
Bacteroides_55	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola plebeius_A
Bacteroides_58	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola sp000432735
Bacteroides_59	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola sp000434735
Bacteroides_60	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola sp000436795
Bacteroides_61	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola sp002161565
Bacteroides_63	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola sp002493165
Bacteroides_65	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola sp900066455
Bacteroides_66	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola sp900540105
Bacteroides_67	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola sp900541515
Bacteroides_68	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola sp900542985
Bacteroides_71	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola sp900546095
Bacteroides_76	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola sp900551645

Group_name	Lin_Phylo
Bacteroides_1	is close to an SGB not present in the SGB.Oct22 database
Bacteroides_4	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_caccae
Bacteroides_5	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_cellulosilyticus
Bacteroides_6	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_clarus
Bacteroides_9	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_eggerthii
Bacteroides_10	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_faecis
Bacteroides_11	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_finegoldii
Bacteroides_12	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_fluxus
Bacteroides_13	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_fragilis
Bacteroides_14	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_fragilis
Bacteroides_17	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_intestinalis
Bacteroides_18	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Mediterranea;s__Mediterranea_massiliensis
Bacteroides_19	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_ndongoniae
Bacteroides_21	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_nordii
Bacteroides_23	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_ovatus
Bacteroides_25	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_salyersiae
Bacteroides_27	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_sp_AM10_21B
Bacteroides_33	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_intestinalis
Bacteroides_38	is close to an SGB not present in the SGB.Oct22 database
Bacteroides_40	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_stercorirosoris
Bacteroides_41	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_stercoris
Bacteroides_42	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_thetaiotaomicron
Bacteroides_43	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Mediterranea;s__Mediterranea_massiliensis
Bacteroides_44	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_uniformis
Bacteroides_45	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_xylanisolvens
Bacteroides_47	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola_barnesiae
Bacteroides_48	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola_coprocola
Bacteroides_49	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola_coprophilus
Bacteroides_50	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola_vulgatus
Bacteroides_52	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola_massiliensis
Bacteroides_53	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_mediterraneensis
Bacteroides_54	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola_plebeius

Bacteroides_55	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola_plebeius
Bacteroides_58	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__GGB1387;s__GGB1387_SGB1895
Bacteroides_59	is close to an SGB not present in the SGB.Oct22 database
Bacteroides_60	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__GGB1387;s__GGB1387_SGB1894
Bacteroides_61	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_gallinaceum
Bacteroides_63	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola_SGB1810
Bacteroides_65	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_mediterraneensis
Bacteroides_66	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_caecicola
Bacteroides_67	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola_faecicola
Bacteroides_68	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__GGB1385;s__GGB1385_SGB1890
Bacteroides_71	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__GGB1392;s__GGB1392_SGB1913
Bacteroides_76	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola_SGB1898

Group_name	Lin_GTDB
Bacteroides_1	
Bacteroides_4	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides caccae
Bacteroides_5	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides cellulosilyticus
Bacteroides_6	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides clarus
Bacteroides_9	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides eggerthii
Bacteroides_10	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides faecis
Bacteroides_11	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides finegoldii
Bacteroides_12	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides fluxus
Bacteroides_13	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides fragilis
Bacteroides_14	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides hominis
Bacteroides_17	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides intestinalis
Bacteroides_18	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides massiliensis
Bacteroides_19	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides ndongoniae
Bacteroides_21	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides nordii
Bacteroides_23	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides ovatus
Bacteroides_25	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides salyersiae
Bacteroides_27	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides propionicigenes
Bacteroides_33	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides intestinalis_A
Bacteroides_38	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides sp902362375
Bacteroides_40	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides stercorirosoris
Bacteroides_41	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides stercoris
Bacteroides_42	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides thetaiotaomicron
Bacteroides_43	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides pullorum
Bacteroides_44	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides uniformis
Bacteroides_45	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides xylanisolvens
Bacteroides_47	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola barnesiae
Bacteroides_48	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola coprocola
Bacteroides_49	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola coprophilus
Bacteroides_50	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola vulgatus
Bacteroides_52	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola massiliensis
Bacteroides_53	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola mediterraneensis
Bacteroides_54	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola plebeius
Bacteroides_55	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola plebeius_A
Bacteroides_58	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola sp000432735
Bacteroides_59	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola sp000434735
Bacteroides_60	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola sp000436795
Bacteroides_61	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola caecigallinarum
Bacteroides_63	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola faecalis
Bacteroides_65	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola fibrisolvens
Bacteroides_66	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola gallinarum
Bacteroides_67	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola faecicola
Bacteroides_68	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola sp900542985
Bacteroides_71	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola excrementipullorum
Bacteroides_76	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola sp900551645

2.2 Classification overview Prevotellaceae

Group_name	UHGG_rep	SGB_ID	%	#down	#avail	%GTDB	Type Strain
Prevotella_1	MGYG000000760	SGB1517	50	10	48	50	added (Hoylesella buccalis)
Prevotella_10	MGYG000003353	SGB1504	100	10	24	100	added (DSM 18810)
Prevotella_12	MGYG000003691	SGB1559	100	10	22	100	added
Prevotella_13	MGYG000000272	SGB1617	80	10	162	80	added (Segatella hominis)
Prevotella_14	MGYG000004629	SGB1560	100	10	19	100	added
Prevotella_16	MGYG000001164	SGB1585	100	10	89	100	added (Leyella lascolaii)
Prevotella_22	MGYG000001097	SGB1662	100	10	123	100	added
Prevotella_23	MGYG000002394	SGB1588	100	10	259	100	added
Prevotella_25	MGYG000000779	SGB1684	100	10	42	100	not available
Prevotella_26	MGYG000003680	SGB1590	100	10	18	100	not available
Prevotella_27	MGYG000001763	SGB1680	100	10	434	100	added (Leyella stercorea DSM 18206)
Prevotella_28	MGYG000003912	SGB1697	100	10	60	100	not available
Prevotella_29	MGYG000000695	SGB1614	100	10	84	100	not available
Prevotella_30	MGYG000001042	SGB1592	100	10	44	100	not available
Prevotella_31	MGYG000004698	SGB1589	90	10	19	90	not available
Prevotella_32	MGYG000002133	SGB1658	100	10	211	100	not available
Prevotella_33	MGYG000002653	SGB1606	100	9	9	100	added (Xylanibacter brevis P6B11)
Prevotella_35	MGYG000004371	SGB1666	100	10	251	100	added (Prevotella dentalis DSM 3688)
Prevotella_36	MGYG000004390	SGB1677	100	10	459	100	not available
Prevotella_38	MGYG000000631	SGB1701	100	10	174	100	not available
Prevotella_39	MGYG000004632	SGB1705	100	10	94	100	added
Prevotella_4	MGYG000003374	SGB1557	100	10	59	100	added (DSM 20514)
Prevotella_40	MGYG000001720	SGB1689	100	10	10	100	added (Hallella mizrahii)
Prevotella_43	MGYG000001866	SGB1699	100	10	903	100	not available
Prevotella_45	MGYG000002108	SGB1676	100	10	179	100	added (Prevotella merdae)
Prevotella_46	MGYG000001770	SGB1613	100	10	525	100	not available
Prevotella_50	MGYG000001001	SGB1687	100	10	39	100	not available
Prevotella_51	MGYG000002558	SGB1600	100	10	94	100	added
Prevotella_52	MGYG000001131	SGB1692	100	10	14	100	not available
Prevotella_53	MGYG000000534	SGB1668	100	10	310	100	not available
Prevotella_54	MGYG000001933	SGB1671	100	10	23	100	not available
Prevotella_55	MGYG000002080	SGB1638	100	10	28	100	not available
Prevotella_58	MGYG000000644	SGB1636	100	10	76	100	not available
Prevotella_59	MGYG000000707	SGB1672	100	10	85	100	not available
Prevotella_6	MGYG000000289	SGB1531	100	10	74	100	added (Segatella buccae)
Prevotella_62	MGYG000000553	SGB1623	100	9	9	100	not available
Prevotella_63	MGYG000003814	SGB1581	100	10	28	100	not available
Prevotella_64	MGYG000004491	SGB1663	100	10	55	100	not available
Prevotella_68	MGYG000004640	SGB1596	100	10	12	100	not available
Prevotella_70	MGYG000001975	SGB1673	100	10	46	100	not available
Prevotella_71	MGYG000000701	SGB1667	100	10	72	100	not available
Prevotella_73	MGYG000004758	SGB1675	100	10	11	100	not available
Prevotella_74	MGYG000002737	SGB1586	100	10	12	100	not available
Prevotella_75	MGYG000003461	SGB1670	100	10	15	100	not available
Prevotella_76	MGYG000001806	SGB1601	100	9	9	100	not available
Prevotella_77	MGYG000001040	SGB1593	100	10	13	100	not available
Prevotella_78	MGYG000003346	SGB1595	100	10	10	100	not available
Prevotella_79	MGYG000001034	SGB1587	100	10	52	100	not available
Prevotella_8	MGYG000003812	SGB1502	100	9	9	100	added (Hallella colorans)
Prevotella_82	MGYG000001056	SGB1612	100	10	15	100	not available
Prevotella_83	MGYG000004456	SGB1635	100	10	20	100	not available
Prevotella_85	MGYG000002293	SGB1626	90	10	2479	90	added (Segatella copri DSM 18205)
Prevotella_86	MGYG000001240	SGB1664	100	10	12	100	not available
Prevotella_89	MGYG000002603	SGB1653	100	10	12	100	not available

Prevotella_9	MGYG000003697	SGB1644	70	10	383	70	not available
Prevotella_97	MGYG000000215	SGB1679	100	10	405	100	included
Prevotella_98	MGYG000003794	SGB1507	90	10	20	90	added (Hoylesella timonensis but not on SGB?)

6 Supplementary Material

Group_name	Species_UHGG	Species_Phylo	Species_GTDB
Prevotella_1		Prevotella buccalis	Prevotella buccalis
Prevotella_10	Prevotella corporis	Prevotella corporis	Prevotella corporis
Prevotella_12	Prevotella disiens	Prevotella disiens	Prevotella disiens
Prevotella_13	Prevotella hominis	Prevotella hominis	Prevotella hominis
Prevotella_14	Prevotella intermedia	Prevotella intermedia	Prevotella intermedia
Prevotella_16	Prevotella lascolaii	Prevotella lascolaii	Prevotella lascolaii
Prevotella_22	Prevotella pectinovora	Prevotella pectinovora	Prevotella pectinovora
Prevotella_23	Prevotella rara	Prevotella marseillensis	Prevotella rara
Prevotella_25	Prevotella sp000431975	GGB1256 SGB1684	Prevotella sp000431975
Prevotella_26	Prevotella sp000433175	GGB1221 SGB1590	Prevotella sp000433175
Prevotella_27	Prevotella sp000434975	Prevotella stercorea	Prevotella sp000434975
Prevotella_28	Prevotella sp000435635	GGB1264 SGB1697	Prevotella avicola
Prevotella_29	Prevotella sp000436035	Prevotella copri clade F	Prevotella sp000436035
Prevotella_30	Prevotella sp000436595	Marseilla massiliensis	Prevotella sp000436595
Prevotella_31	Prevotella sp000436695	Prevotella SGB1589	Prevotella sp000436695
Prevotella_32	Prevotella sp000436915	GGB1239 SGB1658	Prevotella sp000436915
Prevotella_33	Prevotella sp002251295	Prevotella brevis	Prevotella sp002251295
Prevotella_35	Prevotella sp002251385	Prevotella dentalis	Prevotella sp002251385
Prevotella_36	Prevotella sp002265625	Prevotella sp 885	Prevotella sp002265625
Prevotella_38	Prevotella sp002299635	GGB12972 SGB1701	Prevotella sp002299635
Prevotella_39	Prevotella sp002300055	Hallella faecis	Prevotella faecis
Prevotella_4	Prevotella bivia	Prevotella bivia	Prevotella bivia
Prevotella_40	Prevotella sp002350355	Prevotella mizrahii	Prevotella mizrahii
Prevotella_43	Prevotella sp003447235	GGB1266 SGB1699	Prevotella sp003447235
Prevotella_45	Prevotella sp900290275	Prevotella sp Marseille P4119	Prevotella merdae
Prevotella_46	Prevotella sp900313215	Prevotella copri clade B	Prevotella sp900313215
Prevotella_50	Prevotella sp900540375	Prevotella SGB1687	Prevotella sp900540375
Prevotella_51	Prevotella sp900540415	Marseilla massiliensis	Prevotella intestinigallinarum
Prevotella_52	Prevotella sp900543585	GGB1260 SGB1692	Prevotella sp900543585
Prevotella_53	Prevotella sp900543975	GGB1247 SGB1668	Prevotella sp900543975
Prevotella_54	Prevotella sp900544495	Prevotella SGB1671	Prevotella sp900544495
Prevotella_55	Prevotella sp900544825	Prevotella SGB1638	Prevotella sp900544825
Prevotella_58	Prevotella sp900546535	Prevotella copri clade D	Prevotella sp900546535
Prevotella_59	Prevotella sp900546575	Prevotella SGB1672	Prevotella sp900546575
Prevotella_6	Prevotella buccae	Prevotella buccae	Prevotella buccae
Prevotella_62	Prevotella sp900548535	Prevotella copri clade I	Prevotella sp900548535
Prevotella_63	Prevotella sp900548585	GGB1215 SGB1581	Prevotella sp900548585
Prevotella_64	Prevotella sp900548745	GGB1243 SGB1663	Prevotella sp900548745
Prevotella_68	Prevotella sp900551055	GGB74447 SGB1596	Prevotella sp900551055
Prevotella_70	Prevotella sp900552515	Prevotella stercorea	Prevotella sp900552515
Prevotella_71	Prevotella sp900552675	GGB1246 SGB1667	Prevotella sp900552675
Prevotella_73	Prevotella sp900553155	Prevotella SGB1675	Prevotella sp900553155
Prevotella_74	Prevotella sp900553465	GGB1219 SGB1586	Prevotella sp900553465
Prevotella_75	Prevotella sp900553595	GGB1249 SGB1670	Prevotella sp900553595
Prevotella_76	Prevotella sp900553965	GGB1228 SGB1601	Prevotella sp900553965
Prevotella_77	Prevotella sp900554045	GGB1223 SGB1593	Prevotella stercoripullorum
Prevotella_78	Prevotella sp900554545	GGB74447 SGB1595	Prevotella sp900554545
Prevotella_79	Prevotella sp900554695	GGB1219 SGB1587	Prevotella sp900554695
Prevotella_8	Prevotella colorans	Prevotella colorans	Prevotella colorans
Prevotella_82	Prevotella sp900556395	Prevotella copri clade H	Prevotella sp900556395
Prevotella_83	Prevotella sp900556795	Prevotella SGB1635	Prevotella sp900556795
Prevotella_85	Prevotella sp900557255	Prevotella copri clade A	Prevotella sp015074785
Prevotella_86	Prevotella sp900557405	GGB1243 SGB1664	Prevotella sp900557405
Prevotella_89	Prevotella sp900767615	Prevotella copri clade G	Prevotella sp900767615
Prevotella_9	Prevotella copri_A	Prevotella copri clade C	Prevotella copri_A
Prevotella_97	Prevotella stercorea	Prevotella stercorea	Prevotella stercorea
Prevotella_98	Prevotella timonensis	Prevotella timonensis	Prevotella timonensis

Group_name	Lin_UHGG
Prevotella_1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__
Prevotella_10	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella corporis
Prevotella_12	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella disiens
Prevotella_13	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella hominis
Prevotella_14	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella intermedia
Prevotella_16	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella lascolaii
Prevotella_22	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella pectinovora
Prevotella_23	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella rara
Prevotella_25	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp000431975
Prevotella_26	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp000433175
Prevotella_27	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp000434975
Prevotella_28	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp000435635
Prevotella_29	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp000436035
Prevotella_30	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp000436595
Prevotella_31	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp000436695
Prevotella_32	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp000436915
Prevotella_33	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp002251295
Prevotella_35	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp002251385
Prevotella_36	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp002265625
Prevotella_38	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp002299635
Prevotella_39	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp002300055
Prevotella_4	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella bivia
Prevotella_40	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp002350355
Prevotella_43	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp003447235
Prevotella_45	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900290275
Prevotella_46	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900313215
Prevotella_50	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900540375
Prevotella_51	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900540415
Prevotella_52	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900543585
Prevotella_53	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900543975
Prevotella_54	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900544495
Prevotella_55	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900544825
Prevotella_58	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900546535
Prevotella_59	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900546575
Prevotella_6	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella buccae
Prevotella_62	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900548535
Prevotella_63	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900548585
Prevotella_64	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900548745
Prevotella_68	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900551055
Prevotella_70	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900552515
Prevotella_71	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900552675
Prevotella_73	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900553155
Prevotella_74	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900553465
Prevotella_75	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900553595
Prevotella_76	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900553965
Prevotella_77	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900554045
Prevotella_78	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900554545
Prevotella_79	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900554695
Prevotella_8	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella colorans
Prevotella_82	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900556395
Prevotella_83	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900556795
Prevotella_85	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900557255
Prevotella_86	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900557405
Prevotella_89	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900767615
Prevotella_9	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella copri_A
Prevotella_97	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella stercorea
Prevotella_98	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella timonensis

Group_name	Lin_Phylo
Prevotella_1	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_buccalis
Prevotella_10	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_corporis
Prevotella_12	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_disiens
Prevotella_13	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_hominis
Prevotella_14	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_intermedia
Prevotella_16	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_lascolaii
Prevotella_22	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_pectinovora
Prevotella_23	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_marseillensis
Prevotella_25	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB1256;s__GGB1256_SGB1684
Prevotella_26	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB1221;s__GGB1221_SGB1590
Prevotella_27	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_SGB1680
Prevotella_28	d__Bacteria;p__Bacteroidetes;c__CFGB1264;o__OFGB1264;f__FGB1264;g__GGB1264;s__GGB1264_SGB1697
Prevotella_29	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_copri_clade_F
Prevotella_30	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Marseilla;s__Marseilla_massiliensis
Prevotella_31	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_SGB1589
Prevotella_32	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB1239;s__GGB1239_SGB1658
Prevotella_33	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_sp_P3_92
Prevotella_35	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_sp_P4_51
Prevotella_36	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_sp_885
Prevotella_38	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB12972;s__GGB12972_SGB1701
Prevotella_39	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB1270;s__GGB1270_SGB1705
Prevotella_4	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_bivia
Prevotella_40	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_mizrahii
Prevotella_43	d__Bacteria;p__Bacteroidetes;c__CFGB4405;o__OFGB4405;f__FGB4405;g__GGB1266;s__GGB1266_SGB1699
Prevotella_45	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_sp_Marseille_P4119
Prevotella_46	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_copri_clade_B
Prevotella_50	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_SGB1687
Prevotella_51	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Marseilla;s__Marseilla_massiliensis
Prevotella_52	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB1260;s__GGB1260_SGB1692
Prevotella_53	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB1247;s__GGB1247_SGB1668
Prevotella_54	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB33085;s__GGB33085_SGB1671
Prevotella_55	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_SGB1638
Prevotella_58	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_copri_clade_D
Prevotella_59	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB33085;s__GGB33085_SGB1672
Prevotella_6	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_buccae
Prevotella_62	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_copri_clade_I
Prevotella_63	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB1215;s__GGB1215_SGB1581
Prevotella_64	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB1243;s__GGB1243_SGB1663
Prevotella_68	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB74447;s__GGB74447_SGB1596
Prevotella_70	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB33085;s__GGB33085_SGB1673
Prevotella_71	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB1246;s__GGB1246_SGB1667
Prevotella_73	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_SGB1675
Prevotella_74	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB1219;s__GGB1219_SGB1586
Prevotella_75	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB1249;s__GGB1249_SGB1670
Prevotella_76	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB1228;s__GGB1228_SGB1601
Prevotella_77	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB1223;s__GGB1223_SGB1593
Prevotella_78	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB74447;s__GGB74447_SGB1595
Prevotella_79	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB1219;s__GGB1219_SGB1587
Prevotella_8	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_colorans
Prevotella_82	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_copri_clade_H
Prevotella_83	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_clade_D
Prevotella_85	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_copri_clade_A
Prevotella_86	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__GGB1243;s__GGB1243_SGB1664
Prevotella_89	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_copri_clade_G
Prevotella_9	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_copri_clade_C
Prevotella_97	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_stercorea
Prevotella_98	d__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__Prevotella_timonensis

Group_name	Lin_GTDB
Prevotella_1	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella buccalis
Prevotella_10	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella corporis
Prevotella_12	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella disiens
Prevotella_13	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella hominis
Prevotella_14	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella intermedia
Prevotella_16	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella lascolaii
Prevotella_22	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella pectinovora
Prevotella_23	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella rara
Prevotella_25	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp000431975
Prevotella_26	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp000433175
Prevotella_27	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp000434975
Prevotella_28	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella avicola
Prevotella_29	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp000436035
Prevotella_30	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp000436595
Prevotella_31	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp000436695
Prevotella_32	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp000436915
Prevotella_33	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp002251295
Prevotella_35	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp002251385
Prevotella_36	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp002265625
Prevotella_38	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp002299635
Prevotella_39	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella faecis
Prevotella_4	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella bivia
Prevotella_40	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella mizrahii
Prevotella_43	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp003447235
Prevotella_45	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella merdae
Prevotella_46	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900313215
Prevotella_50	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900540375
Prevotella_51	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella intestinigallinarum
Prevotella_52	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900543585
Prevotella_53	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900543975
Prevotella_54	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900544495
Prevotella_55	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900544825
Prevotella_58	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900546535
Prevotella_59	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900546575
Prevotella_6	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella buccae
Prevotella_62	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900548535
Prevotella_63	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900548585
Prevotella_64	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900548745
Prevotella_68	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900551055
Prevotella_70	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900552515
Prevotella_71	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900552675
Prevotella_73	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900553155
Prevotella_74	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900553465
Prevotella_75	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900553595
Prevotella_76	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900553965
Prevotella_77	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella stercoripullorum
Prevotella_78	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900554545
Prevotella_79	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900554695
Prevotella_8	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella colorans
Prevotella_82	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900556395
Prevotella_83	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900556795
Prevotella_85	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp015074785
Prevotella_86	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900557405
Prevotella_89	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp900767615
Prevotella_9	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella copri_A
Prevotella_97	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella stercorea
Prevotella_98	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella timonensis

2.3 Included Metagenomes

NW_ERR2619707	NW_ERR2619708	NW_ERR2619709	NW_ERR2619710	NW_ERR2619711
NW_ERR2619712	NW_ERR2619713	NW_ERR2619714	NW_ERR2619716	NW_ERR2619717
NW_ERR2619718	NW_ERR2619719	NW_ERR2619720	NW_ERR2619722	NW_ERR2619723
NW_ERR2619724	NW_ERR2619725	NW_ERR2619727	NW_ERR2619728	NW_ERR2619729
NW_ERR2619731	NW_ERR2619732	NW_ERR2619733	NW_ERR2619734	NW_ERR2619735
NW_ERR2619736	NW_ERR2619737	NW_ERR2619739	NW_ERR2619740	NW_ERR2619741
NW_ERR2619742	NW_ERR2619743	NW_ERR2619744	NW_ERR2619745	NW_ERR2619746
NW_ERR2619747	NW_ERR2619748	NW_ERR2619749	NW_ERR2619750	NW_ERR2619751
NW_ERR2619752	NW_ERR2619753	NW_ERR2619755	NW_ERR2619757	NW_ERR2619758
NW_ERR2619759	NW_ERR2619760	NW_ERR2619762	NW_ERR2619763	NW_SRR1761667
NW_SRR1761668	NW_SRR1761669	NW_SRR1761670	NW_SRR1761729	NW_SRR1761730
NW_SRR1761731	NW_SRR1761733	NW_SRR1761756	NW_SRR1761757	NW_SRR1761761
NW_SRR1761762	NW_SRR1761766	NW_SRR1761767	NW_SRR1761769	NW_SRR1761770
NW_SRR1761774	NW_SRR1927149	NW_SRR1929408	NW_SRR1929484	NW_SRR1929563
NW_SRR1930121	NW_SRR1930122	NW_SRR1930123	NW_SRR1930133	NW_SRR1930136
NW_SRR1930138	NW_SRR1930140	NW_SRR1930141	NW_SRR1930142	NW_SRR1930143
NW_SRR1930144	NW_SRR1930145	NW_SRR1930149	NW_SRR1930177	NW_SRR1930179
NW_SRR1930187	NW_SRR1930244	NW_SRR2188436	NW_SRR2192754	NW_SRR2192758
NW_SRR2192803	NW_SRR2222398	NW_SRR2222496	NW_SRR2222510	NW_SRR2222805
NW_SRR2222914	NW_SRR2225611	NW_SRR2225747	NW_SRR2225802	NW_SRR2225824
NW_SRR2226815	NW_SRR2226885	NW_SRR2226949	NW_SRR2227356	NW_SRR2227561
NW_SRR2227637	NW_SRR2227725	NW_SRR2227750	NW_SRR2228151	NW_SRR2228184
NW_SRR2228302	NW_SRR2231151	NW_SRR2231290	NW_SRR2231325	NW_SRR2231336
NW_SRR2231363	NW_SRR2231375	NW_SRR2231780	NW_SRR2231785	NW_SRR2231808
NW_SRR2231833	NW_SRR2236790	NW_SRR2244374	NW_SRR2245578	NW_SRR2245604
NW_SRR2247121	NW_SRR2249246	NW_SRR2249440	NW_SRR2250199	NW_SRR2250248
NW_SRR2250408	NW_SRR2250481	NW_SRR2250514	NW_SRR2250547	NW_SRR2251251
NW_SRR2251637	NW_SRR2937620	NW_SRR2937621	NW_SRR2937622	NW_SRR2938189
NW_SRR2938190	NW_SRR2938191	NW_SRR2938192	NW_SRR2938193	NW_SRR2938194
NW_SRR2938195	NW_SRR2938196	NW_SRR2938197	NW_SRR2938199	NW_SRR2938200
NW_SRR2938201	NW_SRR2938202	NW_SRR2938208	NW_SRR2938209	NW_SRR2938210
NW_SRR2938211	NW_SRR2938212	NW_SRR2938213	NW_SRR2938221	NW_SRR2938222
NW_SRR2938223	NW_SRR2938226	NW_SRR2938227	NW_SRR2938229	NW_SRR2938230
NW_SRR2938233	NW_SRR2938234	NW_SRR2938236	NW_SRR2938238	NW_SRR2938243
NW_SRR2938244	NW_SRR2938245	NW_SRR2938246	NW_SRR2938247	NW_SRR2938248
NW_SRR2938253	NW_SRR2938254	NW_SRR2938255	NW_SRR2938256	NW_SRR2938259
NW_SRR2938260	NW_SRR2938262	NW_SRR2938263	NW_SRR2938309	NW_SRR2938310
NW_SRR3992958	NW_SRR3992959	NW_SRR3992960	NW_SRR3992961	NW_SRR3992962
NW_SRR3992963	NW_SRR3992964	NW_SRR3992965	NW_SRR3992966	NW_SRR3992967
NW_SRR3992991	NW_SRR3992992	NW_SRR3992993	NW_SRR3992994	NW_SRR3992995
NW_SRR3992996	NW_SRR3992997	NW_SRR3992998	NW_SRR3993000	NW_SRR3993001
NW_SRR3993002	NW_SRR3993003	NW_SRR3993004	NW_SRR3993005	NW_SRR3993006
NW_SRR3993007	NW_SRR3993008	NW_SRR3993009	NW_SRR3993025	NW_SRR3993026
NW_SRR3993027	NW_SRR3993028	NW_SRR3993029	NW_SRR3993030	NW_SRR3993031
NW_SRR3993033	NW_SRR3993034	NW_SRR3993035	NW_SRR3993036	NW_SRR3993037
NW_SRR3993038	NW_SRR3993039	NW_SRR3993040	NW_SRR3993041	NW_SRR3993042
NW_SRR3993044	NW_SRR3993045	NW_SRR3993046	NW_SRR3993047	NW_SRR5763445
NW_SRR5763446	NW_SRR5763447	NW_SRR5763448	NW_SRR5763449	NW_SRR5763450
NW_SRR5763451	NW_SRR5763452	NW_SRR5763454	NW_SRR5763455	NW_SRR5763456
NW_SRR5763457	NW_SRR5763458	NW_SRR5763459	NW_SRR5763460	NW_SRR5763461
NW_SRR5763463	NW_SRR5763464	NW_SRR5763465	NW_SRR5763466	NW_SRR5763467
NW_SRR5763468	NW_SRR5763471	NW_SRR5763474	NW_SRR5763475	NW_SRR5763476
NW_SRR5763477	NW_SRR5763478	NW_SRR5763479	NW_SRR5763482	NW_SRR5763484
NW_SRR6060195	NW_SRR6060196	NW_SRR6060197	NW_SRR6060198	NW_SRR6060199
NW_SRR6060200	NW_SRR6060201	NW_SRR6060202	NW_SRR6060203	NW_SRR6060204
NW_SRR6060205	NW_SRR6060206	NW_SRR6060207	NW_SRR6060208	NW_SRR6060209
NW_SRR6060210	NW_SRR6060211	NW_SRR6060212	NW_SRR6060213	NW_SRR6060214
NW_SRR6060215	NW_SRR6060216	NW_SRR6060217	NW_SRR6060218	NW_SRR7658580
NW_SRR7658581	NW_SRR7658583	NW_SRR7658584	NW_SRR7658585	NW_SRR7658588
NW_SRR7658589	NW_SRR7658596	NW_SRR7658599	NW_SRR7658603	NW_SRR7658604
NW_SRR7658606	NW_SRR7658609	NW_SRR7658610	NW_SRR7658611	NW_SRR7658615
NW_SRR7658616	NW_SRR7658618	NW_SRR7658619	NW_SRR7658620	NW_SRR7658624
NW_SRR7658626	NW_SRR7658627	NW_SRR7658628	NW_SRR7658632	NW_SRR7658633
NW_SRR7658635	NW_SRR7658637	NW_SRR7658644	NW_SRR7658646	NW_SRR7658651
NW_SRR7658653	NW_SRR7658657	NW_SRR7658663	NW_SRR7658665	NW_SRR7658667
NW_SRR7658668	NW_SRR7658672	NW_SRR7658676	NW_SRR7658677	NW_SRR7658678

NW_SRR7658680	NW_SRR7658682	NW_SRR7658683	NW_SRR7658684	NW_SRR7658686
NW_SRR7658687	NW_SRR7658688	NW_SRR7658690	NW_SRR8180446	NW_SRR8180447
NW_SRR8180448	NW_SRR8180449	NW_SRR8180450	NW_SRR8784353	NW_SRR8784356
NW_SRR8784358	NW_SRR8784360	NW_SRR8784362	NW_SRR8784364	NW_SRR8784368
NW_SRR8784370	NW_SRR8784374	NW_SRR8784377	NW_SRR8784378	NW_SRR8784379
NW_SRR8784381	NW_SRR8784383	NW_SRR8784384	NW_SRR8784387	NW_SRR8784390
NW_SRR8784391	NW_SRR8784395	NW_SRR8791365	NW_SRR8791366	NW_SRR8791367
NW_SRR8791370	NW_SRR8791371	NW_SRR8791375	NW_SRR8791376	NW_SRR8791382
NW_SRR8791383	NW_SRR8791384	NW_SRR8791385	NW_SRR8791386	NW_SRR8791387
NW_SRR8791391	NW_SRR8791392	NW_SRR8791396	NW_SRR8791397	NW_SRR8791398
NW_SRR8791400	NW_SRR8791401	NW_SRR8791404	NW_SRR8791405	NW_SRR8791407
NW_SRR8803176	NW_SRR8803177	NW_SRR8803178	NW_SRR8803179	NW_SRR8803181
NW_SRR8803182	NW_SRR8803185	NW_SRR8803187	NW_SRR8803188	NW_SRR8803189
NW_SRR8803192	NW_SRR8803193	NW_SRR8803194	NW_SRR8803196	NW_SRR8803197
NW_SRR8803198	NW_SRR8803199	NW_SRR8803206	NW_SRR8803207	NW_SRR8803209
NW_SRR8803213	NW_SRR8803214	NW_SRR8803215	NW_SRR8803216	NW_SRR8803218
NW_SRR8803222	NW_SRR8803224	NW_SRR8803225	NW_SRR8803226	NW_SRR8803227
NW_SRR8803229	NW_SRR8803234	NW_SRR8803236	NW_SRR8803237	NW_SRR8803239
NW_SRR8803240	NW_SRR9108942	NW_SRR9108944	NW_SRR9108945	NW_SRR9108946
NW_SRR9108947	NW_SRR9108948	NW_SRR9108949	NW_SRR9108950	NW_SRR9108951
NW_SRR9108952	NW_SRR9108953	NW_SRR9108954	NW_SRR9108955	NW_SRR9108956
NW_SRR9108957	NW_SRR9108958	NW_SRR9108959	NW_SRR9108960	NW_SRR9108961
NW_SRR9108962	NW_SRR9108963	NW_SRR9108964	NW_SRR9108965	NW_SRR9108966
NW_SRR9108967	NW_SRR9108968	NW_SRR9108969	NW_SRR9108970	NW_SRR9108971
NW_SRR9108972	NW_SRR9292735	NW_SRR9292739	NW_SRR9292743	NW_SRR9292783
NW_SRR9292787	NW_SRR9292789	NW_SRR9292820	NW_SRR9292825	NW_SRR9292826
NW_SRR9292828	NW_SRR9292876	NW_SRR9292880	NW_SRR9292881	NW_SRR9292885
NW_SRR9292891	NW_SRR9292904	NW_SRR9292906	NW_SRR9292910	NW_SRR9292984
NW_SRR9293064	NW_SRR9293066	NW_SRR9293070	NW_SRR9293076	NW_SRR9293110
NW_SRR9293114	NW_SRR9293116	NW_SRR9293118	NW_SRR9293120	NW_SRR9293122
NW_SRR9293128	NW_SRR9293218	NW_SRR9293223	NW_SRR9293227	NW_SRR9293248
NW_SRR9293253	NW_SRR9293328	NW_SRR9293332	NW_SRR9293345	NW_SRR9293349
NW_SRR9293353	NW_SRR9293357	NW_SRR9293402	NW_SRR9293417	NW_SRR9293421
NW_SRR9293424	NW_SRR9293432	NW_SRR9293436	NW_SRR9293454	NW_SRR9293458
W_ERR1018185	W_ERR1018188	W_ERR1018189	W_ERR1018190	W_ERR1018191
W_ERR1018193	W_ERR1018195	W_ERR1018196	W_ERR1136792	W_ERR1136793
W_ERR1136794	W_ERR1136795	W_ERR1136796	W_ERR1136797	W_ERR1136798
W_ERR1137230	W_ERR1293500	W_ERR1293548	W_ERR1293741	W_ERR1293817
W_ERR1293861	W_ERR1293877	W_ERR1293881	W_ERR1293917	W_ERR1297682
W_ERR1297746	W_ERR1297762	W_ERR1297798	W_ERR1297832	W_ERR1297845
W_ERR1297846	W_ERR1297850	W_ERR1398075	W_ERR1398078	W_ERR1398129
W_ERR1398153	W_ERR1398173	W_ERR1398206	W_ERR1398214	W_ERR1398242
W_ERR1727454	W_ERR1727562	W_ERR1727704	W_ERR1727736	W_ERR1727744
W_ERR1727752	W_ERR1727976	W_ERR1728278	W_ERR1776124	W_ERR1776125
W_ERR1776126	W_ERR1776127	W_ERR1776128	W_ERR1776129	W_ERR1776130
W_ERR1776132	W_ERR1912991	W_ERR1913025	W_ERR1913040	W_ERR1913044
W_ERR1913060	W_ERR1913084	W_ERR1913100	W_ERR1913116	W_ERR2017804
W_ERR2017805	W_ERR2017806	W_ERR2017807	W_ERR2017808	W_ERR2017809
W_ERR2017810	W_ERR2017814	W_ERR209924	W_ERR209931	W_ERR209936
W_ERR209943	W_ERR209948	W_ERR209953	W_ERR209961	W_ERR209969
W_ERR209971	W_ERR209977	W_ERR209979	W_ERR209985	W_ERR209987
W_ERR209989	W_ERR209991	W_ERR209993	W_ERR2162217	W_ERR2162218
W_ERR2162219	W_ERR2162220	W_ERR2162221	W_ERR2162222	W_ERR2162223
W_ERR2162224	W_ERR247139	W_ERR247140	W_ERR247141	W_ERR247142
W_ERR247143	W_ERR247144	W_ERR247145	W_ERR247146	W_ERR260478
W_ERR260479	W_ERR260481	W_ERR262939	W_ERR262941	W_ERR262942
W_ERR262943	W_ERR262944	W_ERR2726404	W_ERR2726405	W_ERR2726407
W_ERR2726408	W_ERR2726409	W_ERR2726410	W_ERR2726419	W_ERR2726420
W_ERR2855801	W_ERR2855822	W_ERR2855866	W_ERR2855878	W_ERR2855881
W_ERR2855918	W_ERR2855940	W_ERR2855950	W_ERR321079	W_ERR321086
W_ERR321091	W_ERR321098	W_ERR321103	W_ERR321108	W_ERR321116
W_ERR321118	W_ERR3404661	W_ERR3404662	W_ERR3404663	W_ERR3404664
W_ERR3404666	W_ERR3404667	W_ERR3404668	W_ERR3404669	W_ERR3957678
W_ERR3957679	W_ERR3957680	W_ERR3957681	W_ERR3957682	W_ERR3957683
W_ERR3957684	W_ERR3957685	W_ERR4086424	W_ERR4088396	W_ERR414512
W_ERR414513	W_ERR414519	W_ERR414526	W_ERR4330026	W_ERR4330027
W_ERR4330028	W_ERR4330029	W_ERR4330030	W_ERR4330031	W_ERR4330032

W_ERR4330033	W_ERR4560729	W_ERR4560738	W_ERR4560753	W_ERR4563248
W_ERR4563253	W_ERR4563263	W_ERR480459	W_ERR480463	W_ERR480471
W_ERR480475	W_ERR480479	W_ERR480480	W_ERR480485	W_ERR480493
W_ERR504979	W_ERR504980	W_ERR504981	W_ERR504985	W_ERR525699
W_ERR525720	W_ERR525847	W_ERR525911	W_ERR525955	W_ERR525967
W_ERR525999	W_ERR526079	W_ERR527046	W_ERR527047	W_ERR527048
W_ERR527050	W_ERR527051	W_ERR527052	W_ERR527055	W_ERR528329
W_ERR6229432	W_ERR6275675	W_ERR6279618	W_ERR6281658	W_ERR6281677
W_ERR6281944	W_ERR6281997	W_ERR6282070	W_ERR636361	W_ERR636383
W_ERR636385	W_ERR636387	W_ERR636405	W_ERR636407	W_ERR636411
W_ERR636413	W_ERR688505	W_ERR688507	W_ERR688515	W_ERR688517
W_ERR688519	W_ERR688520	W_ERR688524	W_ERR688526	W_ERR719242
W_ERR719251	W_ERR719270	W_ERR719275	W_ERR719288	W_ERR719310
W_ERR719312	W_ERR719328	W_ERR866565	W_ERR866566	W_ERR866567
W_ERR866568	W_ERR866585	W_ERR866596	W_ERR866608	W_ERR866609
W_ERR911953	W_ERR911954	W_ERR911955	W_ERR911956	W_ERR912073
W_ERR912074	W_ERR912123	W_ERR912182	W_ERR9707182	W_ERR9707183
W_ERR9707210	W_ERR9707212	W_ERR9708133	W_ERR9708137	W_ERR9708184
W_ERR9708953	W_SRR059413	W_SRR059425	W_SRR059455	W_SRR059853
W_SRR060353	W_SRR060371	W_SRR060443	W_SRR062103	W_SRR1761734
W_SRR1761735	W_SRR1761736	W_SRR1761737	W_SRR1761738	W_SRR1761739
W_SRR1761740	W_SRR1761741	W_SRR1930247	W_SRR1930248	W_SRR1930250
W_SRR1930251	W_SRR1930253	W_SRR1930255	W_SRR1930777	W_SRR1931170
W_SRR2155180	W_SRR2155181	W_SRR2155323	W_SRR2155324	W_SRR2155339
W_SRR2155358	W_SRR2155377	W_SRR2155482	W_SRR2565998	W_SRR2566006
W_SRR2566008	W_SRR2566009	W_SRR2566013	W_SRR2566015	W_SRR2583965
W_SRR2583968	W_SRR2912775	W_SRR2912776	W_SRR2912787	W_SRR2912798
W_SRR2912807	W_SRR2912808	W_SRR2912809	W_SRR2912810	W_SRR3313034
W_SRR3313035	W_SRR3313046	W_SRR3313057	W_SRR3313068	W_SRR3313079
W_SRR3313090	W_SRR3313102	W_SRR341581	W_SRR341582	W_SRR341583
W_SRR341584	W_SRR341585	W_SRR341587	W_SRR341588	W_SRR341654
W_SRR3992955	W_SRR3992956	W_SRR3992957	W_SRR3992968	W_SRR3992969
W_SRR3992970	W_SRR3992971	W_SRR3992972	W_SRR4052023	W_SRR4052024
W_SRR4052025	W_SRR4052026	W_SRR4052027	W_SRR4052028	W_SRR4052043
W_SRR4052044	W_SRR4074259	W_SRR4074260	W_SRR4074287	W_SRR5127398
W_SRR5127400	W_SRR5127416	W_SRR5127474	W_SRR5127486	W_SRR5127493
W_SRR5127529	W_SRR5127537	W_SRR5127615	W_SRR5127642	W_SRR5127665
W_SRR5127698	W_SRR5127723	W_SRR5127770	W_SRR5127777	W_SRR5127802
W_SRR5127808	W_SRR5127828	W_SRR5127838	W_SRR5127861	W_SRR5274015
W_SRR5274022	W_SRR5274023	W_SRR5274032	W_SRR5274058	W_SRR5274073
W_SRR5274079	W_SRR5274085	W_SRR5275476	W_SRR5275477	W_SRR5275478
W_SRR5275479	W_SRR5275480	W_SRR5275482	W_SRR5275483	W_SRR5275484
W_SRR5279258	W_SRR5279263	W_SRR5279270	W_SRR5279276	W_SRR5279282
W_SRR5279287	W_SRR5279288	W_SRR5279308	W_SRR5575408	W_SRR5575413
W_SRR5575599	W_SRR5575605	W_SRR5575606	W_SRR5651391	W_SRR5651392
W_SRR5651393	W_SRR5651394	W_SRR5651398	W_SRR5651400	W_SRR5651401
W_SRR5651402	W_SRR5665072	W_SRR5665075	W_SRR5665076	W_SRR5665078
W_SRR5665080	W_SRR5665134	W_SRR5665135	W_SRR5665139	W_SRR5898908
W_SRR5898909	W_SRR5898910	W_SRR5898911	W_SRR5898912	W_SRR5898913
W_SRR5898914	W_SRR5898915	W_SRR5930493	W_SRR5930495	W_SRR5930521
W_SRR5930522	W_SRR5930526	W_SRR5930527	W_SRR5930533	W_SRR5930534
W_SRR6000869	W_SRR6000870	W_SRR6000871	W_SRR6000873	W_SRR6000874
W_SRR6000877	W_SRR6000878	W_SRR6000893	W_SRR6028395	W_SRR6028625
W_SRR6028626	W_SRR6028627	W_SRR6028628	W_SRR6028630	W_SRR6028631
W_SRR6028633	W_SRR6784403	W_SRR6784404	W_SRR6784405	W_SRR6784406
W_SRR6784408	W_SRR6784409	W_SRR6784410	W_SRR6784411	W_SRR7281047
W_SRR7281050	W_SRR7281051	W_SRR7281052	W_SRR7281063	W_SRR7281065
W_SRR7281066	W_SRR7281068	W_SRR7716244	W_SRR7716245	W_SRR7716246
W_SRR8097415	W_SRR8097416	W_SRR8097417	W_SRR8097418	W_SRR8097419
W_SRR8097420	W_SRR8097422	W_SRR8097423	W_SRR8146937	W_SRR8146951
W_SRR8146952	W_SRR8146956	W_SRR8146957	W_SRR8675870	W_SRR8675871
W_SRR8675872	W_SRR8675873	W_SRR8675874	W_SRR8675875	W_SRR8675876
W_SRR8675877	W_SRR8865574	W_SRR8865575	W_SRR8865579	W_SRR8865586
W_SRR8865591	W_SRR8865594	W_SRR8865597	W_SRR8865600	W_SRR8942350
W_SRR8942352	W_SRR8942362	W_SRR8942363	W_SRR8942366	W_SRR8942374
W_SRR8942375	W_SRR8942378	W_SRR9033718	W_SRR9033720	W_SRR9033722
W_SRR9033738	W_SRR9033742	W_SRR9033748	W_SRR9033749	W_SRR9033754

3. Scripts

3.1 UHGG_download.sh

```
path=$1
family=$2
second=$3
family2=$4
grep -w "${family}" ${path}/base_files/genomes-all_metadata.tsv > ${path}/data/genomes-"${family#g__}".tsv
if [ "$second" = "yes" ]; then
    grep -w "$family2" ${path}/base_files/genomes-all_metadata.tsv >> ${path}/data/genomes-"${family#g__}".tsv
Fi
awk -F '\t' '{print $15}' ${path}/data/genomes-${family#g__}.tsv | sort | uniq > ${path}/data/${family#g__}_species_list.txt
mkdir ${path}/data/Genomes_${family#g__}
species_num=1
file=$(echo ${path}/data/genomes-${family#g__}.tsv)
echo -e 'Genus\tSpecies\tNumber\tnumber_avail\tnumber_down\tLiniage' > ${path}/data/${family#g__}_sample_overview.txt
echo -e 'Sample Name\tGenus\tSpecies\tGenome Type\tAccession Number\tRepresentative\tLiniage' > ${path}/data/${family#g__}_sample_detail.txt
while  IFS= read -r line;
do
    grep -w "$line" $( echo ${path}/data/genomes-${family#g__}.tsv) | sort -k2,2 -k7,7nr -k8,8n | head -n 10 > ${path}/data/tmp.txt
    g=$( echo $line | awk -F ';' '{print $6}' )
    s=$( echo $line | awk -F ';' '{print $7}' )
    num_avail=$(grep -w "$line" $( echo ${path}/data/genomes-${family#g__}.tsv) | wc -l)
    num_down=$(grep -w "$line" $( echo ${path}/data/genomes-${family#g__}.tsv) | head -n 10 | wc -l)
    echo -e ${family#g__}'\t'${s#s__}'\t'${species_num}'\t'${num_avail}'\t'${num_down}'\t'$line >> ${path}/data/${family#g__}_sample_overview.txt
    i=1
    while  IFS= read -r tmp_line;
        do
        acc=$( echo $tmp_line | awk -F ' ' '{print $1} ' )
        g_type=$( echo $tmp_line | awk -F ' ' '{print $2} ' )
        link=$( echo $tmp_line | awk -F ' ' '{print $NF} ' )
        name=${family#g__}_${species_num}_${i}
        rep=$( echo $tmp_line | awk -F ' ' '{print $14} ' )
        lin=$( echo "$tmp_line" | awk -F '\t' '{print $15}' )
        echo -e $name'\t'${g#g__}'\t'${s#s__}'\t'$g_type'\t'$acc'\t'$rep'\t'$lin >> ${path}/data/${family#g__}_sample_detail.txt
        wget $link -O ${path}/data/Genomes_${family#g__}/${name}.gff.gz -q
        ((i++))
    done < ${path}/data/tmp.txt
    ((species_num++))
done < ${path}/data/${family#g__}_species_list.txt

rm ${path}/data/tmp.txt
mkdir ${path}/data/Genomes_${family#g__}/GFF
mkdir ${path}/data/Genomes_${family#g__}/FASTA
cd ${path}/data/Genomes_${family#g__}
for file in *.gz;
do
    gunzip $file
    csplit -s ${file%.gz} %##FASTA%
    mv xx00 ./FASTA/${file%.gff.gz}.fa
done
rm *.gff
```

3.2 filtering.sh

```
path=$1
family=$2

awk -F '\t' '{print $1}' ${path}/data/${family}_sample_detail.txt >>
${path}/data/${family}_sample_list.txt

while  IFS= read -r line; do
    echo "$line" | awk -F '\t' '$5 >= 9 {print $1 "_" $3 } ' >>
    ${path}/data/${family}_sample_filtered.txt
done < ${path}/data/${family}_sample_overview.txt

while  IFS= read -r line; do
   grep $(echo ${line}_ ) ${path}/data/${family}_sample_list.txt >>
   ${path}/data/${family}_sample_list_filtered.txt
done < ${path}/data/${family}_sample_filtered.txt

mkdir ${path}/data/Genomes_${family}/FASTA_filtered

while  IFS= read -r line; do
    mv ${path}/data/Genomes_${family}/FASTA/${line}.fa
    ${path}/data/Genomes_${family}/FASTA_filtered
done < ${path}/data/${family}_sample_list_filtered.txt

mkdir ${path}/data/${family}_sample_files
mv ${path}/data/${family}*.txt ${path}/data/${family}_sample_files
mv ${path}/data/genomes-${family}.tsv ${path}/data/${family}_sample_files
```

3.3 assign_sgbs.sh

```
path=$1
family=$2
database=$3

module load conda
conda activate PhyloPhlAn_3

phylophlan_assign_sgbs -i ${path}/data/Genomes_${family}/FASTA_filtered -o
${path}/results/SGB_1_${family}_${database} -d ${database} --nproc 8 -n 1 --
database_folder ${path}/base_files/phylophlan_databases
```

3.4 GTDB_denovo.sh

```
path=$1
family=$2
module load conda
conda activate gtdbtk-2.4.0
gtdbtk de_novo_wf --genome_dir ${path}/data/Genomes_${family}/FASTA_filtered \
                  --taxa_filter f__Bacteroidaceae \
                  #--gtdbtk_classification_file
${path}/data/short_GTDB_${family}/gtdbtk.bac120.summary.tsv \
                  --bacteria \
                  --outgroup_taxon p__Chloroflexota \
                  --out_dir  ${path}/data/GTDB2_denovo_${family} \
                  -x .fa \
                  --cpus 20
```

3.5 SGB_tabels.py

```
#!/usr/bin/env python
# coding: utf-8
import sys
if len(sys.argv) != 4:
    print("Usage: python script_name.py path family database ")
    sys.exit(1)
path = sys.argv[1]
family = sys.argv[2]
database = sys.argv[3]
import pandas as pd
SGB_names = {}
SGB_species = {}
file=path+'/base_files/mpa_vJun23_CHOCOPhlAnSGB_202403_species.txt'

with open(file, 'r') as file:
    for line in file:
        parts = line.strip().split('\t')
        if len(parts) >= 2:
            key = parts[0]
            value = parts[1].split(',')[0]
            SGB_names[key] = 'd'+value[1:].replace('|',';')
            tmp = value.split('|')
            SGB_species[key] = tmp[-1][3:].replace('_', ' ')

SGB_names['noSGB'] = 'is close to an SGB not present in the SGB.Oct22 database'
SGB_species['noSGB'] = 'no SGB'

samples = []
filename =
path+'/data/'+family+'_sample_files/'+family+'_sample_list_filtered.txt'
with open(filename, 'r') as file:
    for line in file:
        samples.append(line.strip())

cols = [ 'UHGG_ID', 'SGB_ID', 'UHGG_rep', 'Type', 'species_UHGG', 'species_Phylo',
'lin_UHGG', 'lin_Phylo']
SGB_matrix = pd.DataFrame(index=samples, columns=cols )

i = 1
file = path+'/data/'+family+'_sample_files/'+family+'_sample_detail.txt'
with open(file, 'r') as file:
    for line in file:
        if i == 1 :
            i +=1
            continue
        else:
            tmp = line.split('\t')
            if tmp[0] not in samples :
                continue
            SGB_matrix.loc[tmp[0], 'UHGG_ID'] = tmp[4]
            SGB_matrix.loc[tmp[0], 'UHGG_rep'] = tmp[5]
            SGB_matrix.loc[tmp[0], 'Type'] = tmp[3]
            SGB_matrix.loc[tmp[0], 'lin_UHGG'] = tmp[6].strip().replace(' ', '_')
            if len(tmp[2]) == 0 :
                SGB_matrix.loc[tmp[0], 'species_UHGG'] = 'noSpecies'
            else :
                SGB_matrix.loc[tmp[0], 'species_UHGG'] = tmp[2]
```

```
i = 1
file =  path+'/results/'+'SGB_1_'+family+'_'+database+'.tsv'
with open(file, 'r') as file:
    for line in file:
        if i == 1 :
            i +=1
            continue
        else:
            tmp = line.split('\t')
            if len(tmp) < 2:
                tmp = line.split(' ')
                SGB_matrix.loc[tmp[0], 'SGB_ID'] = 'noSGB'
                SGB_matrix.loc[tmp[0], 'species_Phylo'] = 'noSGB'
                SGB_matrix.loc[tmp[0], 'lin_Phylo'] = 'noSGB'
            else:
                lin=tmp[1].split(':')[2]
                SGB_matrix.loc[tmp[0], 'SGB_ID'] = lin.split('|')[-1][3:]
                SGB_matrix.loc[tmp[0], 'species_Phylo'] =
                lin.split('s__')[1].split('|')[0].replace('_',' ')
                SGB_matrix.loc[tmp[0], 'lin_Phylo'] = 'd'+lin[1:].replace('|', ';')
file_path = path+'/data/'+family+'_sample_files/'+family+'_'+database+'_SGB_matrix.tsv'
SGB_matrix.to_csv(file_path, sep='\t', index=True, index_label='Genomes')
print(f"DataFrame successfully saved as a tab-separated file {file_path}.")
cols = [ 'UHGG_rep', 'SGB_ID', '%', '# down', '# avail', 'species_UHGG', 'species_Phylo',
    'lin_UHGG', 'lin_Phylo']
SGB_matrix_overview = pd.DataFrame(columns=cols )
from collections import Counter
sample = 0
i = 1
file =  path+'/data/'+family+'_sample_files/'+family+'_sample_overview.txt'
with open(file, 'r') as file:
    for line in file:
        if sample == 0 :
            sample +=1
            continue
        else:
            tmp = line.split('\t')
            if int(tmp[4]) < 9 :
                continue
            rep = []
            sgb = []
            lin = []
            for s in range(1, int(tmp[4])+1) :
                rep.append(SGB_matrix.loc[samples[sample-1], 'UHGG_rep'])
                sgb.append(SGB_matrix.loc[samples[sample-1], 'SGB_ID'])
                lin.append(SGB_matrix.loc[samples[sample-1], 'lin_Phylo'])
                sample += 1
            count_sgb=Counter(sgb)
            count_lin=Counter(lin)
            for j in range(0,len(count_sgb)) :
                curr_sgb=count_sgb.most_common(j+1)[j][0]
                   curr_lin=count_lin.most_common(j+1)[j][0]
                   SGB_matrix_overview.loc[i, 'UHGG_rep'] =
                   Counter(rep).most_common(1)[0][0]
                   SGB_matrix_overview.loc[i, 'SGB_ID'] = curr_sgb
                   SGB_matrix_overview.loc[i, '%'] =
                   int((count_sgb.most_common(j+1)[j][1]/int(tmp[4]))*100)
                   SGB_matrix_overview.loc[i, '# down'] = tmp[4]
                   SGB_matrix_overview.loc[i, '# avail'] = tmp[3]
                   SGB_matrix_overview.loc[i, 'species_UHGG'] = tmp[1]
                   SGB_matrix_overview.loc[i, 'lin_UHGG'] = tmp[5].strip()
                   SGB_matrix_overview.loc[i, 'lin_Phylo'] = curr_lin
                   if curr_sgb == 'noSGB' :
                       SGB_matrix_overview.loc[i,'species_Phylo']='no SGB'
                   else :
                       SGB_matrix_overview.loc[i, 'species_Phylo'] =
                       curr_lin.split('s__')[1].split(';')[0].replace('_', ' ')
              i += 1
```

```
file_path=path+'/data/'+family+'_sample_files/'+family+'_'+database+'_SGB_matrix
    _overview.tsv'
SGB_matrix_overview.to_csv(file_path, sep='\t', index=False)
print(f"DataFrame successfully saved as a tab-separated file {file_path}.")

cols = [ 'Group_name', 'UHGG_rep', 'SGB_ID', '%', '# down', '# avail', 'species_UHGG',
    'species_Phylo', 'lin_UHGG', 'lin_Phylo']
SGB_matrix_overview = pd.DataFrame(columns=cols )

from collections import Counter
sample = 0
i = 1
file= path+'/data/'+family+'_sample_files/'+family+'_sample_overview.txt'
with open(file, 'r') as file:
    for line in file:
        if sample == 0 :
            sample +=1
            continue
        else:
            tmp = line.split('\t')
            if int(tmp[4]) < 9 :
                continue
            rep = []
            sgb = []
            lin = []
            for s in range(1, int(tmp[4])+1) :
                rep.append(SGB_matrix.loc[samples[sample-1], 'UHGG_rep'])
                sgb.append(SGB_matrix.loc[samples[sample-1], 'SGB_ID'])
                lin.append(SGB_matrix.loc[samples[sample-1], 'lin_Phylo'])
                sample += 1
            count_sgb=Counter(sgb)
            count_lin=Counter(lin)
            if len(count_sgb) > 1 :
                if count_sgb.most_common(2)[0][1] ==
                   count_sgb.most_common(2)[1][1] :
                    if count_sgb.most_common(2)[0][0] == 'noSGB' :
                        curr_sgb = count_sgb.most_common(2)[1][0]
                        curr_lin = count_lin.most_common(2)[1][0]
                       per=int((count_sgb.most_common(2)[1][1]/int(tmp[4]))*100)
                    else :
                        curr_sgb = count_sgb.most_common(2)[0][0]
                        curr_lin = count_lin.most_common(2)[0][0]
                       per=int((count_sgb.most_common(2)[0][1]/int(tmp[4]))*100)
                else :
                    curr_sgb = count_sgb.most_common(2)[0][0]
                    curr_lin = count_lin.most_common(2)[0][0]
                    per = int((count_sgb.most_common(2)[0][1]/int(tmp[4]))*100)
            else :
                curr_sgb = count_sgb.most_common(1)[0][0]
                curr_lin = count_lin.most_common(1)[0][0]
                per = int((count_sgb.most_common(1)[0][1]/int(tmp[4]))*100)

            SGB_matrix_overview.loc[i, 'Group_name'] = tmp[0]+'_'+tmp[2]
           SGB_matrix_overview.loc[i,'UHGG_rep']=Counter(rep).most_common(1)[0][0]
            SGB_matrix_overview.loc[i, 'SGB_ID'] = curr_sgb
            SGB_matrix_overview.loc[i, '%'] = per
            SGB_matrix_overview.loc[i, '# down'] = tmp[4]
            SGB_matrix_overview.loc[i, '# avail'] = tmp[3]
            SGB_matrix_overview.loc[i, 'species_UHGG'] = tmp[1]
            SGB_matrix_overview.loc[i, 'lin_UHGG'] = tmp[5].strip()
            SGB_matrix_overview.loc[i, 'lin_Phylo'] = SGB_names[curr_sgb]
            if curr_sgb == 'noSGB' :
                SGB_matrix_overview.loc[i, 'species_Phylo'] = 'no SGB'
            else :
                SGB_matrix_overview.loc[i, 'species_Phylo'] =
                   curr_lin.split('s__')[1].split(';')[0].replace('_', ' ')
            i += 1
file_path =
    path+'/data/'+family+'_sample_files/'+family+'_'+database+'_SGB_matrix_summary.tsv
    '
SGB_matrix_overview.to_csv(file_path, sep='\t', index=False)
print(f"DataFrame successfully saved as a tab-separated file {file_path}.")
```

3.6 phylophlan.sh

```
path=$1
family=$2

module load conda
conda activate PhyloPhlAn_3

phylophlan -i FASTA_filtered \
    --input_folder ${path}/data/Genomes_${family}/FASTA_filtered \
    -o $path/data/${family}_phylophlan \
    --databases_folder ${path}/base_files/phylophlan_databases \
    -d phylophlan \
    --diversity medium \
    --fast \
    --configs_folder $path/base_files/ \
    -f supermatrix_aa.cfg \
    --nproc 20 \
    --genome_extension .fa \
    --verbose
```

3.7 curatedMetagenomicData.R

```
1.  library('curatedMetagenomicData')
2.  library(dplyr)
3.  library(DT)
4.  library(stringr)
5.  library(mia)
6.  library(scater)
7.  library(vegan)

8.  sampleMetadata |>
9.    filter(non_westernized == "no") |>
10.   filter(body_site == "stool") |>
11.   filter(age_category == 'adult') |>
12.   datatable(options = list(dom = "t"), extensions = "Responsive")

13. testSamples_W <-
14.   filter(sampleMetadata, non_westernized == "no") |>
15.   filter(body_site == "stool") |>
16.   filter(age_category == 'adult') |>
17.   returnSamples("relative_abundance", rownames = "short")

18. teststudy_W <-
19.   as.data.frame(colData(testSamples_W)) |>
20.   distinct(study_name)

21. teststudyvec_W <- teststudy_W$study_name
22. studydata_W <- data.frame()
23. for (e in teststudyvec_W) {
24.   Samples <-
25.     filter(sampleMetadata, non_westernized == "no") |>
26.     filter(study_name == e) |>
27.     filter(body_site == "stool") |>
28.     filter(age_category == 'adult') |>
29.     filter(body_site == "stool") |>
30.     slice(1:5) |>
31.     returnSamples("relative_abundance", rownames = "short")

32.   tmp <- as.data.frame(colData(Samples))
33.   if (nrow(tmp) == 0) {
34.     next
35.   }

36.   studydata_W <- bind_rows(studydata_W, as.data.frame(colData(Samples)) )
37. }

38. ids_W <- c()
39. r <- 1
40. while (r <= nrow(studydata_W)) {
41.   tmp <- unlist(strsplit(studydata_W[r,20], split = ';'))
42.   tmp <- tmp[1:3]
43.   ids_W <- c(ids_W, tmp)
44.   r <- r+1
45. }
46. ids_no_na_W <- na.omit(ids_W)
47. ids_no_na_W <- as.vector(ids_no_na_W)
48. writeLines(ids_no_na_W, 'sample_W_IDs.txt')
```

```
sampleMetadata |>
  filter(non_westernized == "yes") |>
  filter(body_site == "stool") |>
  filter(age_category == 'adult') |>
  datatable(options = list(dom = "t"), extensions = "Responsive")

testSamples_NW <-
  filter(sampleMetadata, non_westernized == "yes") |>
  filter(body_site == "stool") |>
  filter(age_category == 'adult') |>
  returnSamples("relative_abundance", rownames = "short")

teststudy_NW <-
  as.data.frame(colData(testSamples_NW)) |>
  distinct(study_name)

teststudyvec_NW <- teststudy_NW$study_name

studydata_NW <- data.frame()

for (e in teststudyvec_NW) {
  Samples <-
    filter(sampleMetadata, non_westernized == "yes") |>
    filter(study_name == e) |>
    filter(body_site == "stool") |>
    filter(age_category == 'adult') |>
    filter(body_site == "stool") |>
    slice(1:40) |>
    returnSamples("relative_abundance", rownames = "short")

  tmp <- as.data.frame(colData(Samples))

  if (nrow(tmp) == 0) {
    next
  }

  studydata_NW <- bind_rows(studydata_NW, as.data.frame(colData(Samples)) )
}


ids_NW <- c()
r <- 1
while (r <= nrow(studydata_NW)) {
  tmp <- unlist(strsplit(studydata_NW[r,20], split = ';'))
  tmp <- tmp[1]
  ids_NW <- c(ids_NW, tmp)
  r <- r+1
}

ids_no_na_NW <- na.omit(ids_NW)
ids_no_na_NW <- as.vector(ids_no_na_NW)

  writeLines(ids_no_na_NW, 'NW_sample_IDs.txt')
```

3.8 metaphlan.sh

```
cd $TMPDIR

file=$(ls /lisc/scratch/cube/pejrimovsky/Thesis/data/Metagenomes/W_split_sort* | sed -n ${SLURM_ARRAY_TASK_ID}p)

while  IFS= read -r line;
do
    read_ID=$line
    module load sratoolkit
    prefetch ${read_ID} --max-size u
    fasterq-dump ${read_ID} --threads 8
    rm -rf ${read_ID}
    module load conda
    conda activate metaphlan-4.1.0
    num=$( ls *.fastq| wc -l )
    if [[ $num -ge 2 ]]; then
        metaphlan ${read_ID}_1.fastq,${read_ID}_2.fastq --bowtie2out ${read_ID}.bowtie2.bz2 --nproc 8 --input_type fastq -o W_${read_ID}.txt --bowtie2db /lisc/scratch/mirror/metaphlan/4.1.0
        mv W_${read_ID}.txt /lisc/scratch/cube/pejrimovsky/Thesis/data/Metagenomes/W/
        size=$( du -sh *_1.fastq | awk '{print $1}')
        pat=$(grep @${read_ID} ${read_ID}_1.fastq | wc -l)
        echo $read_ID $size $pat >> /lisc/scratch/cube/pejrimovsky/Thesis/data/Metagenomes/W_size.txt
    else
        metaphlan ${read_ID}.fastq --bowtie2out ${read_ID}.bowtie2.bz2 --nproc 8 --input_type fastq -o W_${read_ID}.txt --bowtie2db /lisc/scratch/mirror/metaphlan/4.1.0
        mv W_${read_ID}.txt /lisc/scratch/cube/pejrimovsky/Thesis/data/Metagenomes/W/
        size=$( du -sh *.fastq | awk '{print $1}')
        pat=$(grep @${read_ID} ${read_ID}.fastq | wc -l)
        echo $read_ID $size $pat >> /lisc/scratch/cube/pejrimovsky/Thesis/data/Metagenomes/W_size.txt
    fi
    rm -rf *
done < $file
```

3.9 Metaphlan.ipynb

```
path='../Thesis/data/Metagenomes/results_1000/merged_abundance_table_species_1000.txt'
species_sample_abundance_matrix = pd.read_csv(path, sep='\t', header=0, index_col=0)

sample_IDs = list(species_sample_abundance_matrix.columns)
all_species = list(species_sample_abundance_matrix.index)
file_path = '../Thesis/data/Metagenomes/species.txt'

with open(file_path, "w") as file:
    # Iterate through each element of the list'
    for item in all_species:
        # Write the element to the file followed by a newline character
        file.write(item + "\n")

W_species_sample_abundance_matrix = species_sample_abundance_matrix.loc[:, species_sample_abundance_matrix.columns.str.startswith('W_')]
NW_species_sample_abundance_matrix = species_sample_abundance_matrix.loc[:, species_sample_abundance_matrix.columns.str.startswith('NW_')]

path='../Thesis/data/Metagenomes/results_1000/W_species_sample_abundance_matrix_1000.tsv'
W_species_sample_abundance_matrix.to_csv(path, sep='\t', index=True, index_label='Species')
path='../Thesis/data/Metagenomes/results_1000/NW_species_sample_abundance_matrix_1000.tsv'
NW_species_sample_abundance_matrix.to_csv(path, sep='\t', index=True, index_label='Species')

# Absence Presence

threshold = 0.5
sample_species_AP_matrix = ( (species_sample_abundance_matrix > threshold) & (species_sample_abundance_matrix < 100)).astype(int)
species_sample_abundance_matrix = sample_species_AP_matrix * species_sample_abundance_matrix

W_sample_species_AP_matrix = sample_species_AP_matrix.loc[:, sample_species_AP_matrix.columns.str.startswith('W_')]
W_species_sample_abundance_matrix = W_sample_species_AP_matrix * W_species_sample_abundance_matrix

NW_sample_species_AP_matrix = sample_species_AP_matrix.loc[:, sample_species_AP_matrix.columns.str.startswith('NW_')]
NW_species_sample_abundance_matrix = NW_sample_species_AP_matrix * NW_species_sample_abundance_matrix

all_filtered_species = []
perc = 0.1
filtered_out_percentages = []
for s in sample_species_AP_matrix.index :
    abund = sample_species_AP_matrix.loc[s].sum()
    if abund > len(sample_species_AP_matrix.sum())*perc :
        all_filtered_species.append(s)
    filtered_out_percentage = (abund / len(sample_species_AP_matrix.sum())) * 100
    if filtered_out_percentage < 2 :
        filtered_out_percentages.append(filtered_out_percentage)  #Store percentage
```

```
filtered_out_percentages.sort()
import matplotlib.pyplot as plt
plt.figure(figsize=(10, 6))
plt.plot(filtered_out_percentages, 'o', color='blue', markersize=3)
plt.ylabel('Percentage of Samples')
plt.title('Species found in % of total samples')
plt.tight_layout()
plt.show()
filtered_out_percentages = []
for s in sample_species_AP_matrix.index :
    abund = sample_species_AP_matrix.loc[s].sum()
    filtered_out_percentage = (abund / len(sample_species_AP_matrix.sum())) * 100
    filtered_out_percentages.append(filtered_out_percentage)  # Store percentage

filtered_out_percentages.sort()
import matplotlib.pyplot as plt
plt.figure(figsize=(10, 6))
plt.plot(filtered_out_percentages, 'o', color='blue', markersize=3)
plt.axhline(y=10, color='red', linestyle='--', linewidth=1.5, label='10%
    threshold')
plt.ylabel('Percentage of Samples')
plt.title('Species found in % of total samples')
plt.tight_layout()
plt.show()
filtered_species_sample_abundance_matrix =
    species_sample_abundance_matrix.loc[all_filtered_species, :]
W_species_sample_abundance_matrix =
    filtered_species_sample_abundance_matrix.loc[:,
    filtered_species_sample_abundance_matrix.columns.str.startswith('W_')]
NW_species_sample_abundance_matrix =
    filtered_species_sample_abundance_matrix.loc[:,
    filtered_species_sample_abundance_matrix.columns.str.startswith('NW_')]
path='../Thesis/data/Metagenomes/results_1000/filtered_merged_abundance_table_spec
    ies_1000.txt'
filtered_species_sample_abundance_matrix.to_csv(path, sep='\t', index=True,
    index_label='Species')
path='../Thesis/data/Metagenomes/results_1000/filtered_W_abundance_table_species_1
    000.txt'
W_species_sample_abundance_matrix.to_csv(path, sep='\t', index=True,
    index_label='Species')
path='../Thesis/data/Metagenomes/results_1000/filtered_NW_abundance_table_species_
    1000.txt'
NW_species_sample_abundance_matrix.to_csv(path, sep='\t', index=True,
    index_label='Species')
threshold = 0.5
filtered_species_sample_AP_matrix = (filtered_species_sample_abundance_matrix >
    threshold).astype(int)
path='../Thesis/data/Metagenomes/results_1000/filtered_species_sample_AP_matrix_10
    00.txt'
filtered_species_sample_AP_matrix.to_csv(path, sep='\t', index=True,
    index_label='Species')

W_filtered_species_sample_AP_matrix = filtered_species_sample_AP_matrix.loc[:,
    filtered_species_sample_AP_matrix.columns.str.startswith('W_')]
NW_filtered_species_sample_AP_matrix = filtered_species_sample_AP_matrix.loc[:,
    filtered_species_sample_AP_matrix.columns.str.startswith('NW_')]

W_B_species_sample_AP_matrix = B_species_sample_AP_matrix.loc[:,
    B_species_sample_AP_matrix.columns.str.startswith('W_')]
NW_B_species_sample_AP_matrix = B_species_sample_AP_matrix.loc[:,
    B_species_sample_AP_matrix.columns.str.startswith('NW_')]
```

```
74. # PB corr
75.
76. Prevotellaceae=['s__Alloprevotella_SGB1466','s__Alloprevotella_sp_oral_taxon_473',
    's__Alloprevotella_tannerae','s__Candidatus_Segatella_albertsiae','s__Candidatus_S
    egatella_caccae','s__Candidatus_Segatella_intestinihominis','s__Candidatus_Segatel
    la_violae','s__GGB1144_SGB1469','s__GGB1145_SGB1471','s__GGB1146_SGB1472','s__GGB1
    147_SGB1473','s__GGB1147_SGB1474','s__GGB1148_SGB1475','s__GGB1149_SGB1476','s__GG
    B1151_SGB1478','s__GGB1153_SGB1481','s__GGB1213_SGB1579','s__GGB1215_SGB1581','s__
    GGB1217_SGB1583','s__GGB1217_SGB1584','s__GGB1219_SGB1586','s__GGB1219_SGB1587','s
    __GGB1221_SGB1590','s__GGB1223_SGB1593','s__GGB1228_SGB1601','s__GGB1239_SGB1657',
    's__GGB1243_SGB1663','s__GGB1243_SGB1664','s__GGB1246_SGB1667','s__GGB1247_SGB1668
    ','s__GGB1248_SGB1669','s__GGB1249_SGB1670','s__GGB1256_SGB1684','s__GGB1277_SGB17
    14','s__GGB12950_SGB20101','s__GGB12950_SGB20102','s__GGB12972_SGB1701','s__GGB129
    72_SGB20130','s__GGB12985_SGB20152','s__GGB12988_SGB27410','s__GGB1341_SGB1799','s
    __GGB18559_SGB27424','s__GGB23938_SGB35682','s__GGB28257_SGB40803','s__GGB74447_SG
    B1595','s__GGB74447_SGB1596','s__GGB82028_SGB1704','s__Hallella_faecis','s__Marsei
    lla_massiliensis','s__Paraprevotella_clara','s__Paraprevotella_xylaniphila','s__Pr
    evotella_amnii','s__Prevotella_bergensis','s__Prevotella_bivia','s__Prevotella_bre
    vis','s__Prevotella_brunnea','s__Prevotella_buccae','s__Prevotella_buccalis','s__P
    revotella_colorans','s__Prevotella_corporis','s__Prevotella_dentalis','s__Prevotel
    la_denticola','s__Prevotella_disiens','s__Prevotella_histicola','s__Prevotella_ihu
    mii','s__Prevotella_intermedia','s__Prevotella_jejuni','s__Prevotella_lascolaii','
    s__Prevotella_marseillensis','s__Prevotellamassilia_timonensis','s__Prevotella_mel
    aninogenica','s__Prevotella_mizrahii','s__Prevotella_nanceiensis','s__Prevotella_n
    igrescens','s__Prevotella_pallens','s__Prevotella_pectinovora','s__Prevotella_SGB1
    589','s__Prevotella_SGB1604','s__Prevotella_SGB1671','s__Prevotella_SGB1672','s__P
    revotella_SGB1675','s__Prevotella_SGB1687','s__Prevotella_SGB1688','s__Prevotella_
    SGB48271','s__Prevotella_sp_885','s__Prevotella_sp_DNF00663','s__Prevotella_sp_Mar
    seille_P4119','s__Prevotella_stercorea','s__Prevotella_timonensis','s__Segatella_b
    rasiliensis','s__Segatella_brunsvicensis','s__Segatella_copri','s__Segatella_homin
    is','s__Segatella_sanihominis','s__Segatella_sinensis','s__Segatella_sinica','s__S
    gatella_sinensis','s__Xylanibacter_rodentium']
77. Bacteroidaceae=['s__Bacteroidaceae_bacterium_14_104','s__Bacteroides_acidifaciens'
    ,'s__Bacteroides_bouchesdurhonensis','s__Bacteroides_caccae','s__Bacteroides_caeci
    cola','s__Bacteroides_caecigallinarum','s__Bacteroides_cellulosilyticus','s__Bacte
    roides_clarus','s__Bacteroides_congonensis','s__Bacteroides_cutis','s__Bacteroides
    _eggerthii','s__Bacteroides_faecalis','s__Bacteroides_faecis','s__Bacteroides_fine
    goldii','s__Bacteroides_fluxus','s__Bacteroides_fragilis','s__Bacteroides_gallinac
    eum','s__Bacteroides_gallinarum','s__Bacteroides_helcogenes','s__Bacteroides_hepar
    inolyticus','s__Bacteroides_intestinalis','s__Bacteroides_mediterraneensis','s__Ba
    cteroides_muris','s__Bacteroides_ndongoniae','s__Bacteroides_nordii','s__Bacteroid
    es_oleiciplenus','s__Bacteroides_ovatus','s__Bacteroides_propionicigenes','s__Bact
    eroides_pyogenes','s__Bacteroides_reticulotermitis','s__Bacteroides_rodentium','s_
    _Bacteroides_salyersiae','s__Bacteroides_SGB1909','s__Bacteroides_SGB1911','s__Bac
    teroides_SGB63885','s__Bacteroides_sp_AF16_49','s__Bacteroides_sp_AF39_11AC','s__B
    acteroides_sp_GM023','s__Bacteroides_sp_Marseille_P3684','s__Bacteroides_sp_OF04_1
    5BH','s__Bacteroides_stercorirosoris','s__Bacteroides_stercoris','s__Bacteroides_t
    hetaiotaomicron','s__Bacteroides_togonis','s__Bacteroides_uniformis','s__Bacteroid
    es_xylanisolvens','s__Bacteroides_zoogleoformans','s__GGB1116_SGB1430','s__GGB1339
    _SGB1796','s__GGB1350_SGB1808','s__GGB1354_SGB1818','s__GGB1355_SGB1819','s__GGB13
    64_SGB1834','s__GGB1379_SGB1880','s__GGB1379_SGB48439','s__GGB1380_SGB1883','s__GG
    B1385_SGB1890','s__GGB1387_SGB1894','s__GGB1387_SGB1895','s__GGB28262_SGB40814','s
    __GGB28265_SGB40817','s__GGB28271_SGB40830','s__GGB47741_SGB65713','s__GGB47742_SG
    B65714','s__GGB50002_SGB70013','s__GGB77003_SGB104903','s__GGB82120_SGB113575','s_
    _Mediterranea_massiliensis','s__Mediterranea_SGB1823','s__Mediterranea_sp_An20','s
    __Phocaeicola_coprocola','s__Phocaeicola_coprophilus','s__Phocaeicola_dorei','s__P
    hocaeicola_faecalis','s__Phocaeicola_faecicola','s__Phocaeicola_massiliensis','s__
    Phocaeicola_plebeius','s__Phocaeicola_salanitronis','s__Phocaeicola_sartorii','s__
    Phocaeicola_SGB1896','s__Phocaeicola_vulgatus']
```

```
78. PB = []
79. P = []
80. B = []
81. for species in Prevotellaceae :
82.     if species in all_filtered_species :
83.         PB.append(species)
84.         P.append(species)
85. for species in Bacteroidaceae :
86.     if species in all_filtered_species :
87.         PB.append(species)
88.         B.append(species)

89. PB_species_sample_abundance_matrix =
    filtered_species_sample_abundance_matrix.loc[PB,:]

90. path='/lisc/scratch/cube/pejrimovsky/Thesis/data/Metagenomes/results_1000/PB_speci
    es_sample_abundance_matrix_1000.txt'
91. PB_species_sample_abundance_matrix.to_csv(path, sep='\t', index=True,
    index_label='Species')

92. PB_AP_matrix = pd.DataFrame(0, index=['Prevotellaceae', 'Bacteroidaceae'],
    columns=sample_IDs )

93. threshold = 0.5
94. PB_species_sample_AP_matrix = (PB_species_sample_abundance_matrix >
    threshold).astype(int)
95. P_species_sample_AP_matrix = PB_species_sample_AP_matrix.loc[P, :]
96. B_species_sample_AP_matrix = PB_species_sample_AP_matrix.loc[B, :]

97. path='/lisc/scratch/cube/pejrimovsky/Thesis/data/Metagenomes/results_1000/PB_speci
    es_sample_AP_matrix_1000.txt'
98. PB_species_sample_AP_matrix.to_csv(path, sep='\t', index=True,
    index_label='Species')

99. for sample in sample_IDs:
100.        if P_species_sample_AP_matrix.loc[:, sample].sum() > 0 :
101.            PB_AP_matrix.loc['Prevotellaceae', sample] = 1
102.        if B_species_sample_AP_matrix.loc[:, sample].sum() > 0 :
103.            PB_AP_matrix.loc['Bacteroidaceae', sample] = 1

104.    path='/lisc/scratch/cube/pejrimovsky/Thesis/data/Metagenomes/results_1000/PB_AP
    _matrix_1000.txt'
105.    PB_AP_matrix.to_csv(path, sep='\t', index=True, index_label='Species')

106.    PB_abundance_matrix = pd.DataFrame(float(0), index=['Prevotellaceae',
    'Bacteroidaceae'], columns=sample_IDs )

107.    for sample in sample_IDs:
108.        PB_abundance_matrix.loc['Prevotellaceae', sample] =
    PB_species_sample_abundance_matrix.loc[P, sample].sum()
109.        PB_abundance_matrix.loc['Bacteroidaceae', sample] =
    PB_species_sample_abundance_matrix.loc[B, sample].sum()

110.    path='/lisc/scratch/cube/pejrimovsky/Thesis/data/Metagenomes/results_1000/PB_ab
    undance_matrix_1000.txt'
111.    PB_abundance_matrix.to_csv(path, sep='\t', index=True, index_label='Species')

112.    PB_stackedbar_abund_matrix = pd.DataFrame(float(0), index=['Prevotellaceae',
    'Bacteroidaceae', 'Other'], columns=sample_IDs )
113.    PB_stackedbar_AP_matrix = pd.DataFrame(float(0), index=['Prevotellaceae',
    'Bacteroidaceae', 'Other'], columns=sample_IDs )
```

```
114.   for sample in sample_IDs:
115.       abund_all = filtered_species_sample_abundance_matrix.loc[: , sample].sum()
116.       abund_P = PB_abundance_matrix.loc['Prevotellaceae', sample].sum()
117.       abund_B = PB_abundance_matrix.loc['Bacteroidaceae', sample].sum()
118.
119.       PB_stackedbar_abund_matrix.loc['Prevotellaceae', sample] = abund_P
120.       PB_stackedbar_abund_matrix.loc['Bacteroidaceae', sample] = abund_B
121.       PB_stackedbar_abund_matrix.loc['Other', sample] = abund_all - abund_P -
    abund_B
122.
123.   for sample in sample_IDs:
124.       abund_all = filtered_species_sample_AP_matrix.loc[: , sample].sum()
125.       abund_P = P_species_sample_AP_matrix.loc[:, sample].sum()
126.       abund_B = B_species_sample_AP_matrix.loc[:, sample].sum()
127.
128.       if abund_all == 0 :
129.           continue
130.       else:
131.           PB_stackedbar_AP_matrix.loc['Prevotellaceae', sample] = abund_P /
    (abund_all /100)
132.           PB_stackedbar_AP_matrix.loc['Bacteroidaceae', sample] = abund_B /
    (abund_all /100)
133.           PB_stackedbar_AP_matrix.loc['Other', sample] = 100 -
    (abund_P/(abund_all/100))- (abund_B/(abund_all/100))
134.
135.   path='/lisc/scratch/cube/pejrimovsky/Thesis/data/Metagenomes/results_1000/PB_st
    ackedbar_abund_matrix_1000.txt'
136.   PB_stackedbar_abund_matrix.to_csv(path, sep='\t', index=True,
    index_label='Species')
137.
138.   path='/lisc/scratch/cube/pejrimovsky/Thesis/data/Metagenomes/results_1000/PB_st
    ackedbar_AP_matrix_1000.txt'
139.   PB_stackedbar_AP_matrix.to_csv(path, sep='\t', index=True,
    index_label='Species')
140.
141.   PB_species_count_matrix = pd.DataFrame(float(0), index=['Prevotellaceae',
    'Bacteroidaceae', 'Other'], columns=sample_IDs )
142.
143.   for sample in sample_IDs:
144.       abund_all = filtered_species_sample_AP_matrix.loc[: , sample].sum()
145.       abund_P = P_species_sample_AP_matrix.loc[:, sample].sum()
146.       abund_B = B_species_sample_AP_matrix.loc[:, sample].sum()
147.
148.       if abund_all == 0 :
149.           continue
150.       else:
151.           PB_species_count_matrix.loc['Prevotellaceae', sample] = abund_P
152.           PB_species_count_matrix.loc['Bacteroidaceae', sample] = abund_B
153.           PB_species_count_matrix.loc['Other', sample] = abund_all
154.
155.   W_PB_species_count_matrix = PB_species_count_matrix.loc[:,
    PB_species_count_matrix.columns.str.startswith('W_')]
156.   NW_PB_species_count_matrix = PB_species_count_matrix.loc[:,
    PB_species_count_matrix.columns.str.startswith('NW_')]
157.
158.   path='/lisc/scratch/cube/pejrimovsky/Thesis/data/Metagenomes/results_1000/W_PB_
    species_count_matrix_1000.txt'
159.   W_PB_species_count_matrix.to_csv(path, sep='\t', index=True,
    index_label='Species')
160.
161.   path='/lisc/scratch/cube/pejrimovsky/Thesis/data/Metagenomes/results_1000/NW_PB
    _species_count_matrix_1000.txt'
   162.       NW_PB_species_count_matrix.to_csv(path, sep='\t', index=True,
       index_label='Species')
```

3.10 Boxplots_with_lines.R

```
library(ggplot2)
library(dplyr)
library(tidyr)
library(tibble)

W <- read.table('W_PB_species_count_matrix_1000.txt', header = TRUE, sep = "\t", row.names = 1)
NW <- read.table('NW_PB_species_count_matrix_1000.txt', header = TRUE, sep = "\t", row.names = 1)
W_long <- W %>%
  rownames_to_column(var = "Group") %>%
  gather(key = "Sample", value = "Value", -Group)
NW_long <- NW %>%
  rownames_to_column(var = "Group") %>%
  gather(key = "Sample", value = "Value", -Group)
W_long$Group <- as.factor(W_long$Group)
NW_long$Group <- as.factor(NW_long$Group)
W_long$Group <- factor(W_long$Group, levels = c("Other",'Prevotellaceae', 'Bacteroidaceae'))
NW_long$Group <- factor(NW_long$Group, levels = c("Other", 'Prevotellaceae', 'Bacteroidaceae'))

group_name <- c('Bacteroidaceae')
for (group_name in c("Other",'Prevotellaceae', 'Bacteroidaceae')) {
  print(group_name)
  W_group <- W_long %>% filter(Group == group_name) %>% pull(Value)
  NW_group <- NW_long %>% filter(Group == group_name) %>% pull(Value)
  W_normality <- shapiro.test(W_group)
  NW_normality <- shapiro.test(NW_group)
  cat("Normality test for W group:\n")
  print(W_normality)
  cat("Normality test for NW group:\n")
  print(NW_normality)
  variance_test <- var.test(W_group, NW_group)
  cat("Variance test result:\n")
  print(variance_test)
  wilcox_test_result <- wilcox.test(W_group, NW_group)
  print(wilcox_test_result)
  cat('')
}

ggplot(data_long, aes(x = interaction(CategoryGroup), y = Count, fill = Group)) +
  geom_boxplot() +
  geom_jitter(width = 0.1, alpha = 0.3) +
  scale_x_discrete(labels = c("", "All Species", "", "Prevotellaceae", "", "Bacteroidaceae", "")) +
  scale_fill_manual(values = c("Non-Westernized" = "maroon", "Westernized" = "navy")) +
  labs(x = "", y = "Count", fill = "Group") +
  theme(
    axis.text.x = element_text(angle = 0, hjust = 1, size = 12),
    plot.margin = unit(c(1, 1, 1, 1), "cm")
  ) +
  geom_vline(xintercept = c(2.5, 4.5), linetype = "dashed", color = "grey")
```

3.11 Correlation_plots.R

```
library(reshape2)
library(ggplot2)
library(dplyr)
a <- read.table('filtered_merged_abundance_table_species_1000.txt', header = TRUE, sep = "\t", row.names = 1)

# Transpose and clean data
data <- as.data.frame(t(a))
constant_columns <- sapply(data, function(x) sd(x) == 0)
data_cleaned <- data[, !constant_columns]

# Calculate correlation matrix
cor_matrix <- cor(data_cleaned, method = "spearman") # pearson spearman

# Melt the correlation matrix
melted_cor_matrix <- melt(cor_matrix)

# Ensure unique pairs and remove self-correlations
melted_cor_matrix <- melted_cor_matrix %>%
  filter(Var1 != Var2) %>%  # Remove self-correlations
  mutate(
    Var1 = as.character(Var1),  # Convert factors to characters
    Var2 = as.character(Var2),
    ordered_pair = ifelse(Var1 < Var2, paste(Var1, Var2, sep = " - "), paste(Var2, Var1, sep = " - "))
  ) %>%
  distinct(ordered_pair, .keep_all = TRUE)  # Keep only unique pairs

# Get top 20 positively correlated pairs
top_pos_cor <- melted_cor_matrix %>%
  arrange(desc(value)) %>%
  slice(1:10)

# Get top 20 negatively correlated pairs
top_neg_cor <- melted_cor_matrix %>%
  arrange(value) %>%
  slice(1:10)

# Combine the top positive and negative correlations
top_cor <- bind_rows(top_pos_cor, top_neg_cor)

# Add a column to indicate positive or negative correlation
top_cor <- top_cor %>%
  mutate(correlation_type = ifelse(value > 0, "Positive", "Negative"))

# Plot the top correlations using a bar plot
plot <- ggplot(data = top_cor, aes(x = reorder(ordered_pair, value), y = value, fill = correlation_type)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = c("Positive" = "maroon", "Negative" = "navy")) +
  coord_flip() +
  theme_minimal() +
  labs(x = "Species Pairs", y = "Pearson Correlation",
       title = "Correlation filtered Species Abundance ") +
  theme(axis.text.y = element_text(size = 20),
        axis.text.x = element_text(size = 20))
print(plot)
```

3.12 Network.R

```
library(NetCoMi)

species_count <- cor_matrix

net_pears <- netConstruct(species_count,
                          dataType = "correlation" ,
                          #measure = "spearman",
                          #normMethod = "clr",
                          #zeroMethod = "multRepl",
                          sparsMethod = "threshold",
                          thresh = 0.3,
                          verbose = 3)

props_pears <- netAnalyze(net_pears, clustMethod = "cluster_fast_greedy")

plot(props_pears,
     layout = "spring",
     nodeColor = "cluster",
     nodeSize = "eigenvector",
     repulsion = 0.7,
     rmSingles = TRUE,
     labelScale = FALSE,
     labelFont = 1,
     cexLabels = 0.75,
     nodeSizeSpread = 3,
     cexNodes = 2,
     hubBorderCol = "darkgray",
     title1 = "Network on Species level with Spearman Correlation",
     showTitle = TRUE,
     cexTitle = 2.3)

legend(0.7, 1.2, cex = 1, title = "estimated correlation:",
       legend = c("+","-"), lty = 1, lwd = 3, col = c("#009900","red"),
       bty = "n", horiz = TRUE, y.intersp = 0.5)
```

3.13 DBSCAN_filter.ipynb

```
import numpy as np
import pandas as pd
from sklearn import metrics
from sklearn.cluster import DBSCAN
from sklearn.neighbors import NearestNeighbors
from matplotlib import pyplot as plt
from sklearn.preprocessing import StandardScaler, normalize

path =
'../Thesis/data/Metagenomes/results_1000/filtered_merged_abundance_table_species_1
000.txt'
df = pd.read_csv(path, sep='\t', header=0, index_col=0)
df = df.T
```

```
12. import numpy as np
13. import pandas as pd
14. from scipy.stats import spearmanr
15. from sklearn.preprocessing import StandardScaler

16. # Rank transform the data (Spearman requires ranked data)
17. ranked_df = df.rank(axis=0)

18. # Scaling the ranked data
19. scaler = StandardScaler()
20. scaled_ranked_df = scaler.fit_transform(ranked_df)

21. # Compute Spearman correlation matrix
22. n = scaled_ranked_df.shape[1]
23. corr_matrix = np.zeros((n, n))

24. for i in range(n):
25.     for j in range(n):
26.         corr_matrix[i, j], _ = spearmanr(scaled_ranked_df[:, i],
    scaled_ranked_df[:, j])

27. # Transform correlation matrix to similarity matrix
28. sim_matrix = (corr_matrix + 1) / 2

29. # Convert similarity matrix to distance matrix
30. distance_matrix = 1 - sim_matrix
31. np.fill_diagonal(distance_matrix, 0)  # Ensure diagonal is zero

32. k=4
33. neighbors = NearestNeighbors(n_neighbors=k, metric='precomputed')
34. neighbors_fit = neighbors.fit(distance_matrix)
35. distances, indices = neighbors_fit.kneighbors(distance_matrix)
36. distances = np.sort(distances, axis=0)
37. distances = distances[:,k-1]
38. plt.plot(distances)

39. import numpy as np
40. import matplotlib.pyplot as plt
41. from scipy.stats import gaussian_kde

42. # Create the histogram
43. plt.figure(figsize=(8, 5))
44. count, bins, _ = plt.hist(distances, bins=20, edgecolor='black', alpha=0.7,
    density=True)

45. # Compute and plot the KDE curve
46. kde = gaussian_kde(distances)
47. x = np.linspace(min(distances), max(distances), 100)
48. plt.plot(x, kde(x), 'g-', linewidth=2, label="KDE Curve")

49. # Labels, title, and legend
50. plt.xlabel("Distance to 3rd Nearest Neighbor")
51. plt.ylabel("Density")
52. plt.title("Histogram of 3rd Nearest Neighbor Distances with KDE Curve")
53. plt.legend()

54. plt.tight_layout()
55. plt.savefig("../Plots/DBSCAN_kdistance_kde.png", dpi=300, bbox_inches='tight')
56. plt.show()
```

```
for e in np.round(np.arange(0.2, 1, 0.025),3) :
    db = DBSCAN(eps=e, min_samples=k, metric='precomputed').fit(distance_matrix)
    labels = db.labels_
    # Number of clusters in labels, ignoring noise if present.
    n_clusters_ = len(set(labels)) - (1 if -1 in labels else 0)
    n_noise_ = list(labels).count(-1)
    print(f'Epsilon: {e}')
    print("Estimated number of clusters: %d" % n_clusters_)
    print("Estimated number of noise points: %d" % n_noise_)
    print("")
db = DBSCAN(eps=0.35, min_samples=k, metric='precomputed').fit(distance_matrix)
cluster_labels = db.labels_
clusters_dict = {}
species_names = df.columns

for species, cluster in zip(species_names, cluster_labels):
    if cluster not in clusters_dict:
        clusters_dict[cluster] = []
    clusters_dict[cluster].append(species)
path = '../Thesis/data/Metagenomes/comb_1000/merged_taxa_list.txt'
fam_dict = {}
with open(path, "r") as file:
    for line in file:
        tax = line.strip().split('|')
        if len(tax) == 7 :
            if tax[4] not in fam_dict.keys() :
                fam_dict[tax[4]] = [tax[6]]
            elif tax[6] not in fam_dict[tax[4]] :
                fam_dict[tax[4]].append(tax[6])
family_counts = {cluster: {family: 0 for family in fam_dict} for cluster in
clusters_dict}
for cluster, species_list in clusters_dict.items():
    for family, family_species in fam_dict.items():
        family_counts[cluster][family] = sum(1 for species in species_list if
species in family_species)

family_counts_df = pd.DataFrame(family_counts).T
column_sums = family_counts_df.sum()
family_counts_df = family_counts_df.loc[:, column_sums != 0]

import matplotlib.cm as cm#
from itertools import cycle

num_families = len(family_counts_df.columns)
colors_cycle = cycle(cm.get_cmap('tab20').colors + cm.get_cmap('tab20b').colors +
cm.get_cmap('tab20c').colors)
colors = [next(colors_cycle) for _ in range(num_families)]

fig, ax = plt.subplots(figsize=(12, 6))
  family_counts_df.plot(kind='bar', stacked=True, ax=ax, color=colors)

  ax.legend(title="Families", bbox_to_anchor=(1.05, 1), loc='upper left')
  ax.set_title("Composition of Clusters by Family")
  ax.set_xlabel("Clusters")
  ax.set_ylabel("Number of Species")
  plt.tight_layout(rect=[0, 0, 0.85, 1])
  plt.show()
```

3.14 HDBSCAN.ipynb

```
path =
'../Thesis/data/Metagenomes/results_1000/filtered_merged_abundance_table_species_1
000.txt'
df = pd.read_csv(path, sep='\t', header=0, index_col=0)
df = df.T
import numpy as np
import pandas as pd
from scipy.stats import spearmanr
from sklearn.preprocessing import StandardScaler

# Rank transform the data (Spearman requires ranked data)
ranked_df = df.rank(axis=0)

# Scaling the ranked data
scaler = StandardScaler()
scaled_ranked_df = scaler.fit_transform(ranked_df)

# Compute Spearman correlation matrix
n = scaled_ranked_df.shape[1]
corr_matrix = np.zeros((n, n))
for i in range(n):
    for j in range(n):
        corr_matrix[i, j], _ = spearmanr(scaled_ranked_df[:, i],
scaled_ranked_df[:, j])

# Transform correlation matrix to similarity matrix
sim_matrix = (corr_matrix + 1) / 2
# sim_matrix = np.clip(sim_matrix, 0, 1)

# Convert similarity matrix to distance matrix
distance_matrix = 1 - sim_matrix
np.fill_diagonal(distance_matrix, 0)  # Ensure diagonal is zero
db = HDBSCAN(min_cluster_size=4, metric='precomputed').fit(distance_matrix)
cluster_labels = db.labels_
clusters_dict = {}
species_names = df.columns
for species, cluster in zip(species_names, cluster_labels):
    if cluster not in clusters_dict:
        clusters_dict[cluster] = []
    clusters_dict[cluster].append(species)

import matplotlib.pyplot as plt
# Create the main plot
fig, ax = plt.subplots(figsize=(12, 6))
family_counts_df.plot(kind='bar', stacked=True, ax=ax, color=colors)

# Create the legend but don't keep it on the main plot
legend = ax.legend(title="Families", bbox_to_anchor=(1.05, 1), loc='upper left')
ax.set_title("Composition of Clusters by Family - HDBScan")
ax.set_xlabel("Clusters")
ax.set_ylabel("Number of Species")

# Remove the legend from the plot before saving
legend.remove()

# Save the figure WITHOUT the legend
plt.tight_layout()
fig.savefig("../Plots/HDBScan_clusters.png", dpi=300, bbox_inches='tight')
```

```
species_columns = df.columns.tolist()
# Run HDBSCAN on the original data and get the true labels
original_clusterer = HDBSCAN(min_cluster_size=4,
    metric='precomputed').fit(distance_matrix)
original_labels = original_clusterer.labels_
# Define number of species to shuffle per sample
num_species_list = [0, 5, 10, 20, 30, 40, 50, 60, 70, 85]
# Number of repetitions per setting
n_repeats = 10
# To store results
results = {num: [] for num in num_species_list}
for num_switch in num_species_list:
    print(f"Shuffling {num_switch} species per sample:")
    for repeat in range(n_repeats):
        # Copy the original dataframe
        permuted_df = df.copy()
        # Shuffle num_switch species **within each sample**
        for idx in permuted_df.index:
            if num_switch > 0:
                species_to_shuffle = np.random.choice(species_columns,
    size=num_switch, replace=False)
                original_values = permuted_df.loc[idx, species_to_shuffle].values
                np.random.shuffle(original_values)
                permuted_df.loc[idx, species_to_shuffle] = original_values
        # Rank transform the data (Spearman requires ranked data)
        ranked_df = permuted_df.rank(axis=0)
        # Scaling the ranked data
        scaler = StandardScaler()
        scaled_ranked_df = scaler.fit_transform(ranked_df)
        # Compute Spearman correlation matrix
        n = scaled_ranked_df.shape[1]
        corr_matrix = np.zeros((n, n))
        for i in range(n):
            for j in range(n):
                corr_matrix[i, j], _ = spearmanr(scaled_ranked_df[:, i],
    scaled_ranked_df[:, j])
        # Transform correlation matrix to similarity matrix
        sim_matrix = (corr_matrix + 1) / 2
        # sim_matrix = np.clip(sim_matrix, 0, 1)
        # Convert similarity matrix to distance matrix
        distance_matrix = 1 - sim_matrix
        np.fill_diagonal(distance_matrix, 0)  # Ensure diagonal is zero
        # Re-cluster the permuted data
        clusterer = HDBSCAN(min_cluster_size=4, metric='precomputed')
        permuted_labels = clusterer.fit_predict(distance_matrix)
        # Compute ARI between original and permuted labels
        if np.unique(permuted_labels).size > 1:
            ari = adjusted_rand_score(original_labels, permuted_labels)
        else:
            ari = 0.0
        results[num_switch].append(ari)
        print(f"  Repeat {repeat + 1}/{n_repeats}: ARI = {ari:.3f}")
# Compute mean and std of ARI for each num_switch
mean_ari = [np.mean(results[num]) for num in num_species_list]
std_ari = [np.std(results[num]) for num in num_species_list]
# Plot the results
 sns.set(style="whitegrid")
 plt.figure(figsize=(10, 6))
 plt.errorbar(num_species_list, mean_ari, yerr=std_ari, fmt='-o', capsize=4)
 plt.title("Average ARI vs. Number of Species Shuffled", fontsize=14)
 plt.xlabel("Number of species shuffled per sample", fontsize=12)
 plt.ylabel("Adjusted Rand Index (ARI)", fontsize=12)
 plt.ylim(-0.05, 1.05)
 plt.tight_layout()
 plt.show()
```

3.15 Spectral_fiolter.ipynb

```
range_n_clusters = [2, 3, 4, 5, 6]
for n_clusters in range_n_clusters:
    fig, (ax1) = plt.subplots(1, 1)
    fig.set_size_inches(5, 4)
    ax1.set_xlim([-0.3, 0.5])
    ax1.set_ylim([0, len(sim_matrix) + (n_clusters + 1) * 10])
    # Initialize the clusterer with n_clusters value and a random generator
    # seed of 10 for reproducibility.
    clustering = SpectralClustering(n_clusters=n_clusters, affinity='precomputed')
    cluster_labels = clustering.fit_predict(sim_matrix)

    silhouette_avg = silhouette_score(distance_matrix, cluster_labels,
metric='precomputed')
    print(
        "For n_clusters =",
        n_clusters,
        "The average silhouette_score is :",
        silhouette_avg,
    )
    # Compute the silhouette scores for each sample
    sample_silhouette_values = silhouette_samples(distance_matrix, cluster_labels,
metric='precomputed')
    y_lower = 10
    for i in range(n_clusters):
        # Aggregate the silhouette scores for samples belonging to
        # cluster i, and sort them
        ith_cluster_silhouette_values = sample_silhouette_values[cluster_labels ==
i]
        ith_cluster_silhouette_values.sort()
        size_cluster_i = ith_cluster_silhouette_values.shape[0]
        y_upper = y_lower + size_cluster_i
        color = cm.nipy_spectral(float(i) / n_clusters)
        ax1.fill_betweenx(
            np.arange(y_lower, y_upper),
            0,
            ith_cluster_silhouette_values,
            facecolor=color,
            edgecolor=color,
            alpha=0.7,
        )
        # Label the silhouette plots with their cluster numbers at the middle
        ax1.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i+1))
        # Compute the new y_lower for next plot
        y_lower = y_upper + 10  # 10 for the 0 samples
    ax1.set_title("The silhouette plot for the various clusters.")
    ax1.set_xlabel("The silhouette coefficient values")
    ax1.set_ylabel("Cluster label")
    # The vertical line for average silhouette score of all the values
    ax1.axvline(x=silhouette_avg, color="red", linestyle="--")
    ax1.set_yticks([])  # Clear the yaxis labels / ticks
    ax1.set_xticks([-0.3, -0.1, 0, 0.1, 0.3, 0.5])

    plt.suptitle("Silhouette analysis for Spectral Clustering with n_clusters =
%d" % n_clusters,
        fontsize=10,
        fontweight="bold",
    )

plt.show()
```

```
clustering = SpectralClustering(n_clusters=3, affinity='precomputed')
cluster_labels = clustering.fit_predict(sim_matrix)

clusters_dict = {}
species_names = df.columns

for species, cluster in zip(species_names, cluster_labels):
    if cluster not in clusters_dict:
        clusters_dict[cluster] = []
    clusters_dict[cluster].append(species)

path = '../Thesis/data/Metagenomes/comb_1000/merged_taxa_list.txt'
fam_dict = {}
with open(path, "r") as file:
    for line in file:
        tax = line.strip().split('|')
        if len(tax) == 7 :
            if tax[4] not in fam_dict.keys() :
                fam_dict[tax[4]] = [tax[6]]
            elif tax[6] not in fam_dict[tax[4]] :
                fam_dict[tax[4]].append(tax[6])

family_counts = {cluster: {family: 0 for family in fam_dict} for cluster in
clusters_dict}

for cluster, species_list in clusters_dict.items():
    for family, family_species in fam_dict.items():
        family_counts[cluster][family] = sum(1 for species in species_list if
species in family_species)

family_counts_df = pd.DataFrame(family_counts).T
column_sums = family_counts_df.sum()
family_counts_df = family_counts_df.loc[:, column_sums != 0]

import matplotlib.cm as cm#
from itertools import cycle

num_families = len(family_counts_df.columns)
colors_cycle = cycle(cm.get_cmap('tab20').colors + cm.get_cmap('tab20b').colors +
cm.get_cmap('tab20c').colors)
colors = [next(colors_cycle) for _ in range(num_families)]

fig, ax = plt.subplots(figsize=(12, 6))
family_counts_df.plot(kind='bar', stacked=True, ax=ax, color=colors)

# Move the legend outside the plot by calling ax.legend() instead of plt.legend()
ax.legend(title="Families", bbox_to_anchor=(1.05, 1), loc='upper left')
ax.set_title("Composition of Clusters by Family")
ax.set_xlabel("Clusters")
ax.set_ylabel("Number of Species")

legend.remove()

plt.tight_layout(rect=[0, 0, 0.85, 1])  # Adjust plot area to fit the legend
plt.savefig("../Plots/Spectral_clusters.png", dpi=300, bbox_inches='tight')
plt.show()
```

3.16 Cazy.sh

```
path='/lisc/scratch/cube/pejrimovsky/Thesis'
family=CAZY

file=$(ls /lisc/scratch/cube/pejrimovsky/Thesis/data/CAZY_sample_files/split* | sed -n ${SLURM_ARRAY_TASK_ID}p)

module load conda
conda activate CAZyme_annotation

while  IFS= read -r line;
do
    run_dbcan ${path}/data/Genomes_CAZY/FAA/${line}.faa protein --tools hmmer --stp_cpu 32 -c ${path}/data/Genomes_CAZY/GFF/${line}.gff --cgc_substrate --out_dir ${path}/data/Genomes_CAZY/CAZY/${line}.PUL.SUB --dia_cpu 32 --hmm_cpu 32 --tf_cpu 32 --db_dir /lisc/user/pejrimovsky/.conda/envs/CAZyme_annotation/db

done < ${file}
```

3.17 CAZY.ipynb

```
1.  import pandas as pd
2.  import os.path
3.  import numpy as np
4.  from scipy.stats import fisher_exact
5.  from statsmodels.stats.multitest import multipletests
6.  import statistics
7.  import statsmodels.api as sm
8.  path='../Thesis/data/CAZY_overview/all_CAZY_overview.txt'
9.  all_CAZY = pd.read_csv(path, sep='\t', names=['Gene_ID','genome_id','HMMER',
    'no1', 'no2', 'no3'], index_col=0, header=None)
10. all_CAZY = all_CAZY.drop('no1', axis=1)
11. all_CAZY = all_CAZY.drop('no2', axis=1)
12. all_CAZY = all_CAZY.drop('no3', axis=1)
13. genomes=[]
14. for s in all_CAZY.index :
15.     tmp = s.split('_')
16.     new=tmp[0]+'_'+tmp[1]+'_'+tmp[2]+'_'
17.     genomes.append(new)
18. genomes=list(set(genomes))
19. len(genomes)
20. species=[]
21. for s in all_CAZY.index :
22.     tmp = s.split('_')
23.     new=tmp[0]+'_'+tmp[1]+'_'
24.     species.append(new)
25. species=list(set(species))
26. len(species)
27. cluster_dict =
    {-1:['s__Phascolarctobacterium_succinatutens','s__GGB1266_SGB1699','s__Eubacterium
    _siraeum','s__Coprococcus_eutactus','s__GGB9758_SGB15368','s__Lachnospira_pectinos
    chiza','s__Ruminococcus_bromii','s__Escherichia_coli','s__Faecalibacterium_sp_HTFF
    ','s__Roseburia_inulinivorans','s__Roseburia_faecis','s__GGB1146_SGB1472','s__GGB3
    304_SGB4367','s__Lachnospira_eligens','s__Methanobrevibacter_smithii','s__Simiaoa_
    sunii','s__GGB3277_SGB4327','s__Butyrivibrio_crossotus','s__Brotolimicola_acetigig
    nens','s__GGB1495_SGB2071','s__Wujia_chipingensis','s__Ruminococcus_bicirculans','
    s__Akkermansia_muciniphila','s__Sutterella_wadsworthensis','s__GGB51647_SGB4348','
    s__GGB3256_SGB4303','s__Dialister_invisus'],
    0:['s__Faecalibacterium_prausnitzii','s__Candidatus_Cibionibacter_quicibialis','s_
    _Dorea_longicatena','s__Anaerostipes_hadrus','s__Blautia_wexlerae','s__Agathobacul
    um_butyriciproducens','s__Oscillibacter_sp_ER4','s__Blautia_massiliensis','s__Euba
    cterium_rectale','s__Oscillibacter_sp_MSJ_31','s__Oscillibacter_valericigenes','s_
    _Mediterraneibacter_faecis','s__Gemmiger_formicilis','s__Holdemanella_porci','s__A
    naerobutyricum_hallii','s__Coprococcus_comes','s__Blautia_obeum','s__Faecalibacter
    ium_SGB15346','s__GGB1627_SGB2230','s__GGB9059_SGB13976','s__GGB1458_SGB2021','s__
    Fusicatenibacter_saccharivorans','s__GGB9635_SGB15106','s__Collinsella_aerofaciens
    ','s__Bifidobacterium_adolescentis','s__Bifidobacterium_longum'],
    1:['s__GGB2734_SGB3677','s__Segatella_sinica','s__Segatella_copri','s__Segatella_b
    rasiliensis','s__Prevotella_stercorea','s__Segatella_sinensis','s__GGB1239_SGB1657
    ','s__Segatella_hominis','s__Prevotella_sp_Marseille_P4119','s__GGB1247_SGB1668','
    s__Prevotella_sp_885','s__GGB4266_SGB5809','s__Segatella_brunsvicensis','s__Segate
    lla_sanihominis','s__Prevotellamassilia_timonensis','s__GGB1148_SGB1475'],
    2:['s__Bacteroides_fragilis','s__Phocaeicola_vulgatus','s__Alistipes_shahii','s__P
    arabacteroides_distasonis','s__Bacteroides_thetaiotaomicron','s__Bacteroides_unifo
    rmis','s__Bacteroides_ovatus','s__Bacteroides_caccae','s__Phocaeicola_dorei','s__B
    arnesiella_intestinihominis','s__Alistipes_putredinis','s__Phocaeicola_massiliensi
    s','s__Phascolarctobacterium_faecium','s__Alistipes_onderdonkii','s__Bacteroides_s
    tercoris','s  Parabacteroides merdae']}
```

```
28. species_ov = pd.read_csv('cazy_list.txt', sep='\t',
    names=['species','species_name','family'], index_col=0, header=None)

29. species_ov.index = ['L' + str(i)[:-1] for i in species_ov.index]

30. species_ov['species_name'] = species_ov['species_name'].apply(
31.     lambda x: 's__' + x.replace(' ', '_'))

32. species_ov['family'] = species_ov['family'].apply(
33.     lambda x: x.split(';')[4] if isinstance(x, str) and len(x.split(';')) > 4 else
    '')

34. species_ov['cluster_id'] = None
35. for cluster_id, species_list in cluster_dict.items():
36.     species_ov.loc[species_ov['species_name'].isin(species_list), 'cluster_id'] =
    cluster_id

37. genomes_CAZY = pd.DataFrame('NA', index=genomes, columns=['GH','GT','PL', 'CE',
    'AA', 'CBM'])
38. CAZYMES=['GH','GT','PL', 'CE', 'AA', 'CBM']
39. for g in genomes:
40.     tmp = all_CAZY.loc[all_CAZY.index.str.startswith(g),'HMMER']
41.     families=[]
42.     for c in tmp :
43.         cs = c.split('+')
44.         for s in cs:
45.             families.append((s.split('(')[0]))
46.     for caz in CAZYMES :
47.         genomes_CAZY.loc[g,caz]=list([item for item in families if
    item.startswith(caz)])

48. genomes_abs_CAZY = pd.DataFrame(0.0, index=genomes, columns=['GH','GT','PL', 'CE',
    'AA', 'CBM'])
49. for genome in genomes_abs_CAZY.index :
50.     for caz in CAZYMES :
51.         genomes_abs_CAZY.loc[genome,caz] = len(genomes_CAZY.loc[genome,caz])

52. genomes_div_CAZY = pd.DataFrame(0.0, index=genomes, columns=['GH','GT','PL', 'CE',
    'AA', 'CBM'])
53. for genome in genomes_abs_CAZY.index :
54.     for caz in CAZYMES :
55.         genomes_div_CAZY.loc[genome,caz] = len(set(genomes_CAZY.loc[genome,caz]))

56. species_abs_CAZY = pd.DataFrame(0.0, index=species, columns=['GH','GT','PL', 'CE',
    'AA', 'CBM'])
57. for s in species:
58.     tmp = genomes_abs_CAZY.loc[genomes_abs_CAZY.index.str.startswith(s),:]
59.     for caz in CAZYMES:
60.         species_abs_CAZY.loc[s,caz] = sum(tmp.loc[:,caz])/10

61. species_div_CAZY = pd.DataFrame(0.0, index=species, columns=['GH','GT','PL', 'CE',
    'AA', 'CBM'])
62. for s in species:
63.     tmp = genomes_div_CAZY.loc[genomes_div_CAZY.index.str.startswith(s),:]
64.     for caz in CAZYMES:
65.         species_div_CAZY.loc[s,caz] = sum(tmp.loc[:,caz])/10
```

```
CAZYMES = ['GH','GT','PL','CE','AA','CBM']

# New table: index = species_name, value = summed CAZy count
species_cazy_sum = pd.DataFrame(columns=['Total_CAZy'])

for idx in species_abs_CAZY.index:
    # Skip if species ID is not present in species_ov
    if idx not in species_ov.index:
        continue

    # Get species name
    species_name = species_ov.loc[idx, 'species_name']

    # Sum CAZy values for this row
    total_cazy = species_abs_CAZY.loc[idx, CAZYMES].sum()

    # Add or accumulate
    if species_name in species_cazy_sum.index:
        species_cazy_sum.loc[species_name, 'Total_CAZy'] += total_cazy
    else:
        species_cazy_sum.loc[species_name, 'Total_CAZy'] = total_cazy

abund_path =
    '../Thesis/data/Metagenomes/results_1000/filtered_merged_abundance_table_species_10
    00.txt'
abund = pd.read_csv(abund_path, sep='\t', index_col=0)

common_species = abund.index.intersection(species_cazy_sum.index)

abund_sub = abund.loc[common_species]
species_cazy_sub = species_cazy_sum.loc[common_species]

 weighted_cazy_loop = pd.DataFrame(
    0.0,
    index=abund_sub.index,
    columns=abund_sub.columns)

for sp in abund_sub.index:
    abund_values = abund_sub.loc[sp]
    total_cazy = species_cazy_sub.loc[sp, 'Total_CAZy']
   weighted_values = abund_values * total_caz
   weighted_cazy_loop.loc[sp] = weighted_values

   weighted_cazy_species = weighted_cazy_loop.sum(axis=1).to_frame(
       name='Weighted_Total_CAZy')

   wc = weighted_cazy_species.join(
       species_ov.set_index('species_name')['cluster_id'],
       how='left')

   from scipy.stats import kruskal
   vals_0  = wc.loc[wc['cluster_id'] == 0,  'Weighted_Total_CAZy']
   vals_1  = wc.loc[wc['cluster_id'] == 1,  'Weighted_Total_CAZy']
   vals_2  = wc.loc[wc['cluster_id'] == 2,  'Weighted_Total_CAZy']
   stat, p = kruskal(vals_0, vals_1, vals_2)
   print(stat, p)
```

```
from scipy.stats import mannwhitneyu
from itertools import combinations
from statsmodels.stats.multitest import multipletests

clusters = {0: vals_0,1: vals_1,2: vals_2}
pvals = []
pairs = []
for (c1, v1), (c2, v2) in combinations(clusters.items(), 2):
    stat, p = mannwhitneyu(v1, v2, alternative='two-sided')
    pvals.append(p)
    pairs.append((c1, c2))
rej, p_adj, _, _ = multipletests(pvals, method='fdr_bh')
for (c1, c2), pa, r in zip(pairs, p_adj, rej):
    print(f"Cluster {c1} vs {c2}: adj p = {pa}, significant = {r}")

vals_0 = wc.loc[wc['cluster_id'] == 0, 'Weighted_Total_CAZy']
vals_1 = wc.loc[wc['cluster_id'] == 1, 'Weighted_Total_CAZy']
vals_2 = wc.loc[wc['cluster_id'] == 2, 'Weighted_Total_CAZy']

import matplotlib.pyplot as plt
import numpy as np

fig, ax = plt.subplots(figsize=(6, 5))
bp = ax.boxplot(
    [vals_0, vals_1, vals_2],
    labels=['Lachno', 'Prevotella', 'Bacteroides'],
    patch_artist=True,   # this is the key line
    showfliers=True)

colors = ['darkgreen', 'darkred', 'darkblue']

for patch, color in zip(bp['boxes'], colors):
    patch.set_facecolor(color)
    patch.set_alpha(0.7)

for element in ['whiskers', 'caps', 'medians']:
    for item in bp[element]:
        item.set_color('black')

ax.set_ylabel('Abundance-weighted Total CAZy')
ax.set_yscale('log')
ax.set_title('Cohort-level abundance-weighted CAZy contribution by cluster')

plt.tight_layout()
plt.show()
```