Norwegian University
of Life Sciences

Master’s Thesis 2022 60 ECTS
Faculty of Chemistry, Biotechnology and Food Science

The Potential Metabolism of Human Milk Oligosaccharides and Mucin in the Infant Intestine by *Bifidobacterium* and *Bacteroides*

Oda Larsen Hamarheim
MSc Biotechnology

The Potential Metabolism of Human Milk Oligosaccharides and Mucin in the Infant Intestine by *Bifidobacterium* and *Bacteroides*

Norwegian University of Life Sciences (NMBU),

Faculty of Chemistry, Biotechnology and Food Science

Acknowledgements

The work presented in this thesis was performed at the Faculty of Chemistry, Biotechnology and Food Science at the Norwegian University of Life Sciences, under the supervision of Professor Knut Rudi and PhD. Morten Nilsen.

First, I would like to thank my main supervisor Knut Rudi for all the support and enthusiasm I have received during the past year. I am grateful that I got to be a part of your group and that you with delight have shared your knowledge with me. A special thank you to my co-supervisor Morten Nilsen, for his never-ending positivity and patience, and for being the most reliable supporter. He has all the credit for my (previously non-existent) RStudio-skills, and I really appreciate all the hours he has spent troubleshooting with me and helping me out. I consider myself lucky to have had them both as my supervisors.

Also, I would like to direct a thank you to Karin C. Lødrup and the PreventADALL project for providing samples for me to analyse, and for including us masters' students as well in their symposium in December. It was great to get to know the project better and to see its extent has been inspiring.

And last but not least, thank you to all the Master Students I got to share this year with and for keeping me company during the endless hours spent writing on the 3rd floor, but also the long lunches and good laughs. You have contributed more than you know in making this year a good one.

Ås, 2022

Oda Larsen Hamarheim

Abstract

Among the most abundant genera in the gut of 6-month-old infants are the *Bacteroides* and *Bifidobacterium* genera. *Bifidobacterium* species are well known to utilize human milk oligosaccharides (HMOs) but can also degrade mucins present in humans' gastrointestinal tract. Mucins are structurally similar to HMOs and are the primary resource for *Bacteroides* species. *Bacteroides* have recently been discovered to degrade HMOs, and *Bifidobacterium* seems not to be the only species to possess this trait. There is currently a knowledge gap related to the common metabolism of these resources and the potential competition between the *Bacteroides* and *Bifidobacterium* genera. This thesis aims to investigate metabolic pathways and glycoside hydrolases that are recognized for HMO- and mucin degradation with the help of metagenomics and proteomics.

A subset of 11 samples were selected from a bigger sample-set of 100 16S rRNA sequenced fecal samples. The subset was split into two groups: samples high in *Bacteroides* and samples high in *Bifidobacterium*. A shotgun analysis was performed to investigate the potential functions present in *Bacteroides* and *Bifidobacterium* and proteome analysis to identify and match proteins to the shotgun data. Accordingly, an SCFA analysis was performed to identify associations between the produced SCFA and metabolic pathways. All necessary intracellular glycoside hydrolases (GHs) for HMO degradation were detected for both genera in the shotgun data, including sialidases, fucosidases, β-galactosidases, and β-hexosaminidases. Two mucin-related GHs were found in the genome of *Bifidobacterium* and not in *Bacteroides*. Sulfatases that may be used to degrade other substrates in human breast milk were identified in the *Bacteroides* genome and not in *Bifidobacterium*. The proteomics revealed the presence of fucosidases and β-hexosaminidases in both genera. However, sialidases were missing for *Bifidobacterium* and *Bacteroides,* whereas the latter also lacked β-galactosidases. There was no correlation between SCFAs and the two genera, but the potential for producing acetate was observed in different metabolic pathways for *Bifidobacterium* and *Bacteroides*. The latter illustrated a potential for producing propionate as well.

The study revealed similar abilities of HMO and mucin degradation between *Bacteroides* and *Bifidobacterium,* although the *Bifidobacterium* genus is likely better adapted and has a broader repertoire of the necessary enzymes for HMO utilization. *Bacteroides* could depend on other factors in human milk to compete with the *Bifidobacterium* species. The potential competition between these genera and the metabolic pathways they may exploit to promote their species' growth, should be further investigated.

Sammendrag

Blant de mest tallrike bakteriene i tarmen til 6 måneder gamle spedbarn er slektene *Bacteroides* og *Bifidobacterium*. *Bifidobacterium* er godt kjent for å benytte seg av oligosakkarider i brystmelk (HMO), men de kan også degradere muciner som finnes i den gastrointestinale trakten hos mennesker. Mucin er veldig like HMO i strukturen, og er hovedressursen til arter i *Bacteroides* slekten. *Bacteroides* har og nylig blitt observert å kunne degradere HMO, slik at *Bifidobacterium* ser ut til å ikke være den eneste slekten som har denne egenskapen. Foreløpig er det et avvik i kunnskapen relatert til deres felles metabolisme for disse ressursene, som fremmer levedyktigheten og gjør at de bosetter seg i tarmen. Målet med denne oppgaven har vært å undersøke metabolske reaksjonsveier og glykosid hydrolaser som er kjent for HMO- og mucin degradering, ved hjelp av metagenomikk og proteomikk.

Fra et prøvesett på 100 16S rRNA sekvenserte fekale prøver, ble 11 prøver valgt ut. De ble delt inn i to grupper: høy *Bacteroides* og høy *Bifidobacterium.* En shotgun analyse ble gjennomført for å undersøke potensielle funksjoner hos *Bacteroides* og *Bifidobacterium,* samt en proteom-analyse for å identifisere og matche proteiner til shotgun-data. I tillegg ble det gjennomført en kortkjedet fettsyre-analyse for å identifisere assosiasjoner mellom de produserte fettsyrene og metabolske reaksjonsveiene. Alle nødvendige intracellulære glykosid-hydrolaser (GH) for HMO degradering ble funnet for begge slekter i shotgun-dataen. Disse inkluderer sialidaser, fukosidaser, β-galaktosidaser og β-heksosaminidaser. To mucin-relaterte GHer ble funnet i genomet til *Bifidobacterium,* men ikke hos *Bacteroides*. Sulfataser, som kan benyttes for degradering av andre substrater i brystmelk ble funnet hos *Bacteroides* og ikke hos *Bifidobacterium*. Proteomikken avslørte tilstedeværelse av fukosidaser og β-heksosaminidaser hos begge slektene. Sialidaser derimot manglet hos begge, og *Bacteroides* manglet også β-galactosidaser. Ingen korrelasjon mellom kortkjedede fettsyrer og bakterieslektene ble observert, men potensialet for acetat-produksjon ble observert i ulike metabolske reaksjonsveier for *Bifidobacterium* og *Bacteroides*. *Bacteroides* illustrerte også et potensiale for propionat-produksjon.

Både *Bacteroides* og *Bifidobacterium* viste gode muligheter for degradering av HMO og mucin i denne studien, men *Bifidobacterium* synes å være bedre tilpasset med et større repertoar av enzymene som er nødvendige i degradering av HMO. *Bacteroides* kan være avhengig av andre faktorer i brystmelk for å kunne konkurrere med *Bifidobacterium*. Metabolske reaksjonsveier og GHer som *Bacteroides* og *Bifidobacterium* benytter seg av burde undersøkes videre, for å få en bredere forståelse av hva det er som fremmer veksten til disse slektene i tarmen til spedbarn.

Abbreviations

ABC transporter	ATP-binding Cassette transporter
ATP	Adenosine Triphosphate
BCFA	Branched-chain fatty acids
CAZyme	Carbohydrate Active Enzyme
CS	Caesarean-section
ddNTP	Dideoxy nucleotide
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide
Fuc	Fucose
Gal	Galactose
GalE	UDP-glucose-4-epimerase
GalK	Galactokinase
GalM	Galactose Mutarotase
GalNAc	N-acetylgalactosamine
GalT	UDP-glucose-hexose-1-phosphate uridylyl transferase
GC	Gas Chromatography
GH	Glycoside Hydrolases
GI	Gastrointestinal
Glc	Glucose
Glc1P	Glucose-1-phosphate
GlcNAc	N-acetylglucosamine
GLNBP	GNB/LNB phosphorylase
GNB	Galacto-N-biose
HMO	Human Milk Oligosaccharides
LacNAc	N-acetyl lactosamine
LC-MS/MS	Liquid Chromatography- tandem mass spectrometry
LNB	Lacto-N-biose
LPSN	List of Prokaryotic names with Standing in Nomenclature
mRNA	Messenger ribonucleic acid
MS	Mass Spectrometry
Neu5Ac	N-acetyl neuraminic acid/sialic acid
NGS	Next-Generation Sequencing
PCR	Polymerase Chain Reaction
PEP	Phosphoenolpyruvate
RNA	Ribonucleic acid
SCFA	Short-chain fatty acids
SDS-PAGE	Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis
VMHD	The Virtual Metabolic Human Database

Table of Contents

1. Introduction 1

1.1. The Gut Microbiota and General Functions 1

1.2. The Infant Gut Microbiota and Colonization.. 2

1.3. Human Milk Oligosaccharides 3

1.3.1. HMO Utilization and Degradation.. 6

1.4. Mucin O-Glycans.. 8

1.4.1. Mucin Utilization and Degradation.. 9

1.5. HMO- and Mucin- Degrading Bacteria 10

1.6. Short-Chain Fatty Acids 12

1.7. Analytical Approaches to Study the Taxonomic and Functional Aspects of the Gut Microbiota.. 13

1.7.1. Bioinformatic Tools 14

1.7.2. Analyzing Proteins 15

1.7.3. Sequencing Methods for Nucleic Acids.. 15

1.7.4. Analyzing Short-Chain Fatty Acid Composition.. 17

1.8. PreventADALL.. 17

1.9.Aim of Thesis.. 18

2. Materials and Methods 19

2.1. Sample Criteria 20

2.2. Short-Chain Fatty Acid Analysis of Samples with Gas Chromatography.. 20

2.3. Extraction of DNA from Fecal Samples 20

2.4. Measuring DNA Concentrations using Qubit.. 21

2.5. Agarose Gel Electrophoresis.. 22

2.6. Preparing Samples for Shotgun Sequencing.. 22

2.7. Proteome Analysis 23

2.8. Bioinformatic Analysis 26

2.8.1. The Shotgun Database 26

2.8.2. Proteome Analysis after Mass Spectrometry 28

2.9. Statistical Analysis 29

3. Results 31

3.1. Sample Selection 31

3.2. Gut Microbiota Composition in Selected Samples 31

3.3. Composition of *Bacteroides* and *Bifidobacterium* species from Shotgun Sequencing ...32

3.4. Extraction and identification of Unique Genes for *Bacteroides* and *Bifidobacterium* ... 34

3.5. Proteome Analysis 34

3.5.1. Calculations and Clustering of Proteins with Perseus 34

3.5.2. Glycoside Hydrolases 37

3.6. KEGG Mapper Reconstruct – Metabolic pathways 38

3.7. Short-Chain Fatty Acid Analysis 41

4. Discussion 43

4.1. The Leloir- and LNB/GNB- Pathways – Enzymes in the Galactose Metabolism.......... 43

4.2. β-Hexosaminidase Activity needed for HMO- and Mucin Degradation 44

4.3. Degrading Lactose 45

4.4. Fucosidases – the Removal of Fucose from HMOs and Mucin 46

4.5. Sialidases – Sialic Acids as Potential Energy Source 46

4.6. Mucin Related Glycoside Hydrolases 46

4.7. Sulfatases – Necessary to an Extent for Growth of Gut Microbes 47

4.8. Short-Chain Fatty Acids 47

4.8.1. Acetate Production through Pyruvate Metabolism 47

4.8.2. Propionate Production through the Succinate Pathway 48

4.8.3. The Difference in Butyrate Levels between Groups High in *Bacteroides* and High in *Bifidobacterium* 49

4.9.Technical Considerations and Future Research 49

4.9.1. Shotgun Analysis 49

4.9.2. The Proteome Analysis 49

5. Conclusion 50

References 52

Appendix A1 – Tables with Sample Information and Measurements 58

Appendix A2 – Figures 61

Appendix B – Protocols 63

Protocol B1 – GC analysis of short-chain fatty acids in fecal samples 63

Protocol B2 – Mass spectrometry – TimsTOF Aurora 64

Appendix C – R Markdown files 65

C1 – Annotation of Bacterial Taxonomy with the HumGut Database 65

C2 – Get *Bifidobacterium* and *Bacteroides* Proteins from all Faa/Fasta Files 66

C3 – Create a FASTA File for eggNOG Mapper Annotation 70

C4 – “Skal_i_eggNOG.faa”-File Annotated in eggNOG Mapper 71

C5 – K-numbers for the Reconstruction of Pathways 74

C6 – dbCAN Annotation for Glycoside Hydrolase Families 74

C7 – Statistical analysis of Short-Chain Fatty Acids 82

1. Introduction

1.1. The Gut Microbiota and General Functions

The human body harbors a vast number of microorganisms: on the skin, from the air we breathe, and the food we digest. The biggest part of the human microbiome – which means all living microorganisms in the human body - is in the intestine, also known as the gut microbiome. The human intestine inhabits approximately 200-1000 different bacterial species, whereas the composition of these species varies between individuals (Forster et al., 2019). The ratio between bacterial cells and human cells in the human body is approximately 1:1 (Sender et al., 2016), and the total number of genes within these bacterial cells is more than 100 times greater than in the entire human gene set (Qin et al., 2010). The bacterial community lives in symbiosis with its host, meaning they live together, and the relationship between bacteria and humans can be either mutualistic, commensalistic, or parasitic. This ecosystem has proven complex, and even with new bioinformatic tools and expanding knowledge, there are many uncertainties about host-microbe interactions and how microbes affect human health. These interactions have been studied for many years, and the knowledge is continually expanding to understand better the numerous factors that impact human health.

The human gut microbiota composition may have inter-individual variations, but the general functions remain primarily the same. The functional roles can vary among the members of the four major bacterial phyla Bacteroidetes, Firmicutes, Actinobacteria, and Proteobacteria (Forster et al., 2019). For example, did the Actinobacteria investigated in Forster et al. (Forster et al., 2019) show enriched functions associated with lipid and carbohydrate metabolism, while Bacteroidetes had iron and sulfur transporter functions as crucial functions. Humans rely on many of the functions completed by the gut microbiota. They ferment foods, fibers, and nutrients that enter the intestine and generate energy for their benefit, but they also produce vitamins and amino acids that are essential to humans (Masi & Stewart, 2022). Another vital function of the gut microbiome is to defend the human body against pathogens that may enter the gastrointestinal (GI) tract. Many of the immune responses against pathogens, recognition and tolerance of antigens and the commensal flora, and food responses occur in the gastrointestinal system (Vighi et al., 2008). Even if there are many variations in the bacterial composition and abundance of species within the gut of healthy people, distinguishing healthy and unhealthy microbiomes is not easy (Eisenstein, 2020). Nevertheless, a healthy gut microbiome includes a diverse microbiota covering each niche in the intestine which works preventive and protects against colonization of other bacterial species, like pathogens.

1.2. The Infant Gut Microbiota and Colonization

New-born infants show great interindividual variations in their gut bacterial composition but become more alike over time. The bacterial composition in the gut converges from an infant gut microbiota to an adult-like microbiota within the first few years of living – and the most significant alterations occur between birth and two years of age (Avershina et al., 2016; Milani et al., 2017). Around three years of age, it is almost impossible to tell the difference between a child and an adult from their gut microbiota composition (Milani et al., 2017). The first colonization is thought to start at birth, but some studies have suggested *in utero* colonization that indicates a presence of microbes in the amniotic fluid, placenta, and meconium (Collado et al., 2016; Perez-Munoz et al., 2017) – but both the prevalence and abundance has been pretty low in these samples.

The mother has the greatest external impact on the gut colonization of the infant, and mother-infant associations have been found in the first six months after delivery (Matamoros et al., 2013). For example, a previous study associated the mother-child-shared *Bacteroides* with vaginal delivery (Nilsen et al., 2021), suggesting vertical transmission of this genus. The initial establishment is influenced by several factors, and the mode of delivery is one of them. Depending on the delivery mode, the prevalence of different bacterial species in the infant gut varies, as well as the diversity. Typically, a lower diversity during the first week of life has been observed in cesarean-section (CS) delivered infants compared to vaginally delivered infants (Rutayisire et al., 2016). They are also less often colonized by *Bifidobacterium* and *Bacteroides* and more often colonized by *Clostridium* and *Staphylococcus* (Milani et al., 2017; Rutayisire et al., 2016). The maternal perineal and vaginal microbes like *Lactobacillus* and *Prevotella* are natural infant intestinal colonizers for vaginally delivered infants (Hoang et al., 2021; Milani et al., 2017). Other normal inhabitants for vaginal delivery are facultative proteobacteria such as *Escherichia coli* and other *Enterobacteriaceae* (Matamoros et al., 2013). Exposure to the mother's vaginal and fecal microbiota is lacking during CS, and environmental bacteria that exist in the hospital and on hospital staff, including human skin are usual colonizers instead.

The stomach and the small intestine only inhabit a few bacterial species because of the acidic pH, but from the ileum located at the end of the small intestine, the pH gets more basic/alkaline together with the number of bacteria. After birth, oxygen depletion is usually facilitated by the early proteobacteria, which in turn creates an anaerobic environment that favors the growth of strict anaerobe species like *Bifidobacterium, Clostridium,* and *Bacteroides* (Rautava et al., 2012). This process is introduced just after birth and shortly after the entire colon is anaerobic. Because of this, the colon has the densest microbial community and only the species that do not need oxygen in their metabolism and generate energy from fermentation settle here (Lin et al., 2014).

Several factors influence the colonization of the human gut beside the delivery mode (figure 1.1). These can, for example, be the gestational age at birth, feeding mode (breast- or formula-fed), geographical location, medical factors (antibiotic treatments), and familial environment (Matamoros et al., 2013; Milani et al., 2017).

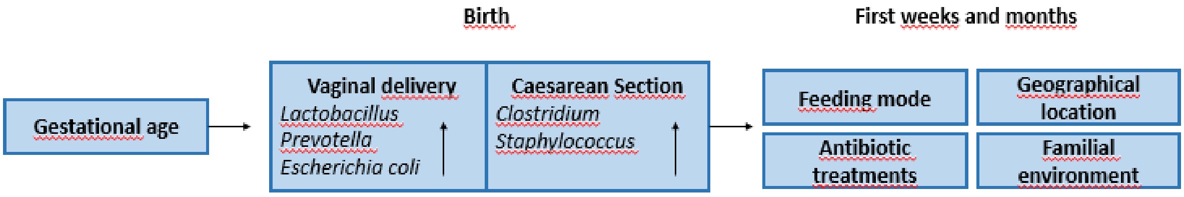

Figure 1.1. General factors that influence the bacterial composition in the gut from birth. The arrows in the "birth" boxes (Vaginal Delivery + Caesarean Section) indicates a usual increase in these genera for these delivery modes. The figure was inspired by Matamoros (Matamoros et al., 2013).

Diet is an important factor from the very beginning since the bacteria's ability to break down different carbohydrates is one of the main factors that determine the gut bacterial composition (Kononova et al., 2021). Human milk is the first diet that humans encounter, and the carbohydrates that are digested are determined by feeding mode, as the carbohydrate composition varies between breastmilk and formula. The exclusive breastfeeding usually lasts for about six months, but variations are great due to cultures, availability, or other social contexts. At six months, children are generally introduced to solid foods as well, making an impact on the gut bacterial succession.

1.3. Human Milk Oligosaccharides

Different lactating stages distinguish human milk into colostrum, transitional and mature milk (Mosca & Gianni, 2017). Colostrum is the milk produced in the first lactating stage in the first days after birth, which primarily benefits the infant gut by transferring immunological factors and growth factors rather than nutrients (Granger et al., 2021; Mosca & Gianni, 2017). The colostrum then transitions into mature breast milk, which is more focused on transferring nutrition for the infant, but also immune factors, growth factors, and vertical transfer of bacteria from the breast milk microbiome (Granger et al., 2021).

Human Milk Oligosaccharides (HMOs) are the third most abundant solid component in breast milk after lactose and lipids. The concentration ranges from approximately 1-10 g/L in mature milk and 15-23 g/L in colostrum (Mosca & Gianni, 2017), but there are some variations between lactating mothers. HMOs are resources that lead to the production of short-chained fatty acids. They are glycan structures that enter the infant intestine from breastmilk without being broken down by the infants' enzymes (Rautava et al., 2012). Instead, these are utilized by bacteria in the infant colon and therefore considered prebiotics and an excellent substrate for bacteria with the correct enzymatic degradation abilities. Because not all bacteria have the necessary enzymes for HMO-degradation, the HMOs contribute to establishing a specialized community (Borewicz et al., 2019).

HMOs consists of five different monosaccharide building blocks, which are D-glucose (Glc), D-galactose (Gal), N-acetylglucosamine (GlcNAc), fucose (Fuc), and sialic acid (N-acetyl neuraminic acid (Neu5Ac)) (figure 1.2) (Borewicz et al., 2019). Lactose, which is Gal-β1,4-Glc, is always found on the reducing end of HMOs and can be elongated with either a β1-3 linkage or β1-4 linkage to two different disaccharides. These are formed by Gal and GlcNAc: Lacto-N-biose (LNB) with a β1-3 linkage creating a type 1 chain, and N-acetyl lactosamine (LacNAc) with a β1-4 linkage creating a type 2 chain (Masi & Stewart, 2022). LNB terminates the chain, and LacNAc can be further elongated (figure 1.2). β1-6 linkages also occur and will create a branched HMO molecule. Further additions may occur if fucose is added with α1-2, α1-3, and α1-4 linkages, and sialic acid with α2-3 and α2-6 linkages (Bode, 2012). These additions term the HMO as fucosylated or sialylated, respectively. A neutral HMO does not have any sialic acid additions, and they make up more than 75% of the total HMO concentration (Wang et al., 2020). Consequently, the sialylated, also known as acidic HMO structures make up about 8-21% of the total HMOs (Ioannou et al., 2021). Together, all these building blocks can create hundreds of different structures, but only 20-25 accounts for more than 95% of the total HMOs (Ioannou et al., 2021).

The structures vary depending on the maternal genotype, which determines the composition and concentration of the separate structures (Borewicz et al., 2019). This means that the oligosaccharide composition is specific to each mother, and it is determined by the activity of two fucosyl-transferases; α1-2-fucosyltransferase FUT2 and α1-3/4-fucosyltransferase FUT3. A variable for HMOs in mothers is the presence or absence of fucose residues on HMOs, which defines their Secretor status and Lewis blood type (Ioannou et al., 2021). Women expressing the *Se* genes are referred to as secretor women, and the gene codes for the activity of FUT2. Lewis positive women express the gene *Le* which codes for the activity of FUT3. *Se*-positive and *Le*-positive women will have high concentrations of HMOs that are either α1-2-fucosylated or α1-4-fucosylated, respectively (Borewicz et al., 2019). Accordingly, *Le*-negative and *Se*-negative women will not have very high levels of these HMOs. Breast milk can be assigned to one of four groups based on the expression of the fucosyltransferases, and these are Le-positive secretors (*Se+Le+*), Le-negative secretors (*Se+Le-),* Le-positive *non*-secretors (*Le+Se-)* and Le-negative *non*-secretors (*Le-Se-*) (Bode, 2012).

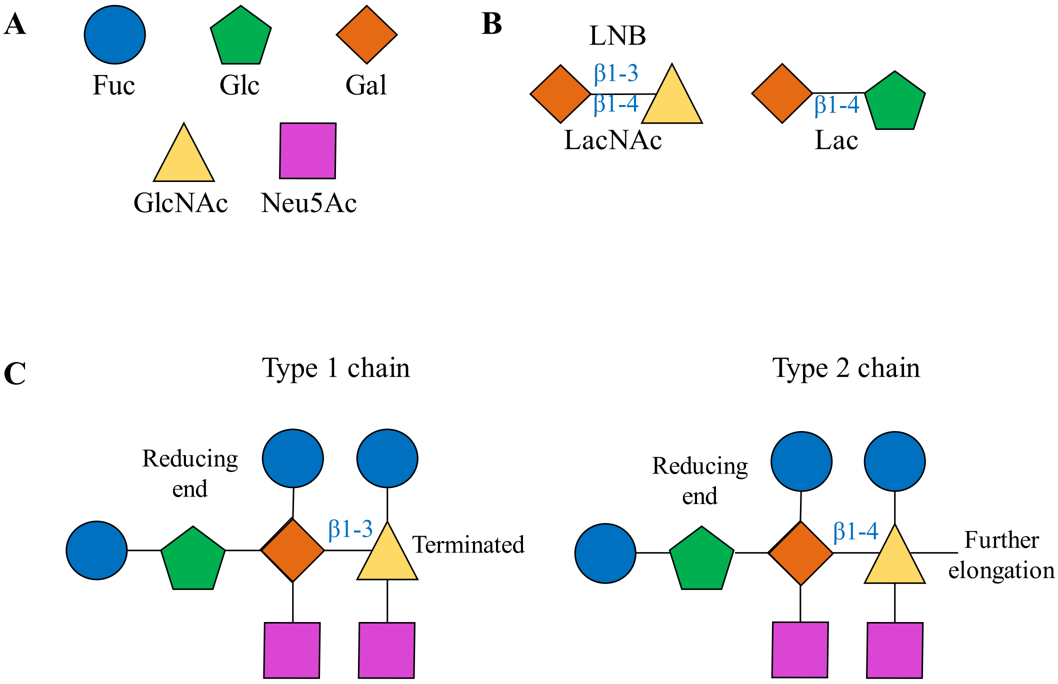

Figure 1.2. Human milk oligosaccharides.
(A) Monosaccharides in HMO. Fuc, fucose; Glc, glucose; Gal, galactose; GlcNAc, N-acetylglucosamine; Neu5Ac, N-acetylneuraminic/sialic acid. **(B)** Disaccharides in HMO. LNB, Lacto-N-Biose; LacNAc, N-acetyl lactosamine; Lac, lactose. Gal and GlcNAc are linked together with a β1-3 (LNB) or β1-4 (GlcNAc) linkage and terminate or extend the chains of HMO. Lactose has a β1-4 linkage between Gal and Glc and is always on the reducing end of HMO. **(C)** Illustrations of a type 1 chain and type 2 chain of HMO, with mono- and disaccharides in (A) and (B). This figure is inspired by and based on (Masi & Stewart, 2022).

Certain inhabitants of the intestine produce enzymes that degrade the linkages in HMO structures, which enables the degraded HMO for utilization by other bacterial species or subspecies. This process is known as cross-feeding, a mechanism that strains of a certain species or even different species use and promote the growth of themselves or others (Turroni et al., 2018).

1.3.1. HMO Utilization and Degradation

HMO degradation in the infant gut is not yet fully understood. However, there are glycosidases, sugar transporters, and glycan-binding proteins necessary for HMO degradation – at least in the well-studied *Bifidobacterium spp*. In extracellular hydrolase-dependent HMO degradation in *Bifidobacterium*, HMOs are hydrolyzed into mono- and disaccharides by secretory glycoside hydrolases and then incorporated inside the cells (Gotoh et al., 2018). Before they are incorporated, the mono- and disaccharides will be available for other bacteria to utilize. Oligosaccharide transporter-dependent degradation is when HMOs are directly imported into the cells by ATP-binding cassette (ABC) transporters (Gotoh et al., 2018). And then, inside the cells, the saccharides are hydrolyzed by intracellular glycosidases.

Glycoside Hydrolases

Numerous studies and researchers have tried to characterize enzymes that degrade HMOs. These enzymes are known as carbohydrate-active enzymes (CAZymes), categorized into different families. A family contains at least one member and is then populated by homologous sequences in the Carbohydrate Active Enzymes database (Drula et al., 2021). Specific glycoside hydrolases (GHs) are known for HMO degradation, and they react with water to break glycosidic linkages (Ioannou et al., 2021). GHs are classified into 171 families to this date (Drula et al., 2021). Substrate specificity can vary in the family, but for GHs, the catalytic mechanism seems to be well conserved (Drula et al., 2021).

Lacto-N-Biose (LNB) is always found on the terminating end of HMO (figure 1.2). Galacto-N-biose (GNB) is structurally similar to LNB but is found in mucins in the mucosal barrier in the intestine in certain core-structures. Both glycans have to be cleaved to make the rest of the substrate available for degradation (Ioannou et al., 2021). The enzyme LNB/GNB phosphorylase (GLNBP, EC 2.4.1.211) cleaves the LNB or GNB, and belong to the GH112 family. Endo-β-N-acetylglucosaminidases (GH18 or GH85) remove glycans from peptides and they may target β1-4 linkages in LacNAc as well as β1-3 linkages of LNB or GNB. Neu5Ac residues, or sialic acids, are freed and cleaved by sialidases/neuraminidases (GH33). α-L-fucosidases (GH29 and GH95), depending on their specificity, cleave the fucose residues of HMOs (Ioannou et al., 2021). Other enzymes involved in HMO degradation are Lacto-N-biosidases (GH20 and GH136) and β-hexosaminidases/β-1,6-N-acetylglucosaminidases (GH20) that release lactose, and β-galactosidases from families GH2 and GH42 including the β-galactosidases from family GH35 (Ioannou et al., 2021; Marcobal et al., 2011). It is the β-galactosidases able to degrade β1,4 linkages that cleave the lactose. The GH1 family seems to not have been characterized in species of the infant gut flora, however, they may target β1,4-linkages of lactose. An overview of the GH families with HMO degrading enzymes and their target substrates can be found in table 1.1.

Table 1.1. Overview of GH families. Illustrates their enzymes and targeted bond that may be involved in HMO and mucin degradation. This table is made with the same information as in (Ioannou et al., 2021) and table 1 in the applicable article and with information from (Tailford et al., 2015).

	GH Family	**Enzyme**	**Target**
HMO related	**GH18**	Endo-β-N-acetylglucosaminidase/ Endoglycosidase	Galβ1-3/4GlcNAc$_2$
	GH112	GNB/LNB phosphorylase	Galβ1-3GlcNAc$_2$/ Galβ1-3GalNAc$_2$
	GH136	Lacto-N-biosidase	GlcNAcβ1-3Gal
	GH1	β-1,4-galactosidase	Galβ1-4-Glc
	GH35	β-galactosidase	Galβ1-4-Glc
Mucin related	**GH89**	α-N-acetylglucosaminidase	GlcNAcα1-4Gal
	GH101	α-N-acetylgalactosaminidase	(Galβ1-3)GalNAc-Ser/Thr
	GH129	α-N-acetylgalactosaminidase	(Galβ1-3)GalNAc-Ser/Thr
HMO and mucin related	**GH95**	α-1,2-L-fucosidase	Fucα1-2Gal
	GH2	β-1,4-galactosidase	Galβ1-4-Glc
	GH42	β-galactosidase	Galβ1-4-Glc + Galβ1-3-Gal + Galβ1-4-Gal + Galβ1-6-Gal
	GH20	Lacto-N-biosidase	GlcNAcβ1-3/6Gal
		β-hexosaminidase/ β-1,6-N-acetylglucosaminidase	
	GH29	α-L-fucosidase	Fucα1-3/4Gal
		α-1,3/1,4-L-fucosidase	Fucα1-3/4GlcNAc
	GH33	2,3-2,6-α-sialidase	Neu5Acα2-3/6Gal + Neu5Acα2-6GlcNAc
	GH85	Endo-β-N-acetylglucosaminidase/ Endoglycosidase	Galβ1-3/4GlcNAc$_2$

Multiple pathways are used to utilize HMO by different bacteria. One of them is the Leloir pathway which consists of four enzymes: galactose mutarotase (GalM); galactokinase (GalK); UDP-glucose-hexose-1-phosphate uridylyltransferase (GalT); and UDP-glucose-4-epimerase (GalE) (Nishimoto & Kitaoka, 2007). Galactose is metabolized to Glucose-1P through this pathway in most organisms and Glucose-1P is then used in the glycolysis. It starts with β-Galactose, which is converted to α-Galactose by GalM (EC 5.1.3.3), and GalK (EC 2.7.1.6) converts the α-Galactose to Galactose-1P. Further, GalT (EC 2.7.7.12) makes the UDP-Galactose and converts UDP-Glucose to Glucose-1P. GalE (EC 5.1.3.2) finishes the pathway by converting UDP-Galactose to UDP-Glucose, which enters the GalT reaction to Glucose-1P (Nishimoto & Kitaoka, 2007). The LNB/GNB pathway is similar to the Leloir pathway, but they have a few differences. The key enzyme is the GLNBP whereas the resulting N-acetylhexosamines are further phosphorylated by NahK (EC 2.7.1.162) (Nishimoto & Kitaoka, 2007). Additionally, the pathway likely uses GalT2 instead of GalT1 due to higher affinity toward GalNAc1P (N-acetyl-galactosamine-1-phosphate) than Gal1P (De Bruyn et al., 2013). It is a more energy-conserving pathway and is mostly used by some *Bifidobacterium* species as the main metabolic pathway for galactose as an energy source (Nishimoto & Kitaoka, 2007).

1.4. Mucin O-Glycans

Throughout the GI tract, mucus covers the surfaces of the epithelial cells, creating a barrier that separates pathogens and other harmful organisms and agents from the epithelial cells, preventing inflammation and colorectal cancer (Luis et al., 2021). Mucus is present in a two-layered system in the colon, where the bacterial density is highest. The inner layer functions as an actual barrier, thick and attached to the epithelium, while the outer layer is looser and is where the commensal bacteria colonize (Raimondi et al., 2021). As mucus is the first barrier that agents and organisms must interact with and diffuse through to access other organs, the risk of disease increases if the mucosal barrier is in some way eliminated or the level of glycosylation is reduced (Bansil & Turner, 2006; Luis et al., 2021). This happens if mucin-degrading species are overrepresented, disassembling the mucin oligosaccharides, resulting in a thin mucosal layer that exposes the epithelial cells (Raimondi et al., 2021).

Mucins are the main component of mucus and are made up of glycoproteins with high levels of fucosylation, which is the reason for the viscous properties of mucus (Bansil & Turner, 2006). They can be membrane-bound or secreted and utilized as a nutrient source for certain bacterial species, such as *Bacteroides* spp. (Luis et al., 2021; Raimondi et al., 2021). Also, their structure is quite similar to HMO. The oligosaccharides in mucin represent ~ 80% of the mucin mass, which are N-acetylgalactosamine (GalNAc), GlcNAc, Gal, Fuc, and Neu5Ac, whereas the latter four are also components of HMO (Raimondi et al., 2021). Mucins are called O-glycans because their oligosaccharide chains are attached to the protein core with an O-glycosidic linkage on the side-chain of serine or threonine (Bansil & Turner, 2006). This protein core makes up the remaining 20% of mucin. There are eight different protein cores (figure 1.3), with threonine or serine always linked to GalNAc, which can be extended by backbone chains. As in HMOs, the backbone consists of either a type 1 (β1-3 linkage) or type 2 (β1-4 linkage) chain and can also be branched (Li & Chai, 2019) (figure 1.3). The backbone is linked to a variable peripheral part that is recognized depending on a person’s blood group antigens and lewis antigens but can also consist of substituents such as sialic acid, fucose or sulfate (Li & Chai, 2019; Luis et al., 2021). Sulfate can be added to the 6-hydroxyl of N-acetyl-D-Glucosamine

(6S-GlcNAc) and the terminal D-galactose (Gal) on the hydroxyl position 3, 4 or 6 (3S-Gal, 4S-Gal and 6S-Gal) (Luis et al., 2021). It caps the glycan so that it is unavailable for further degradation, and mucin-utilizing species are therefore dependent on sulfatases that removes the sulfate.

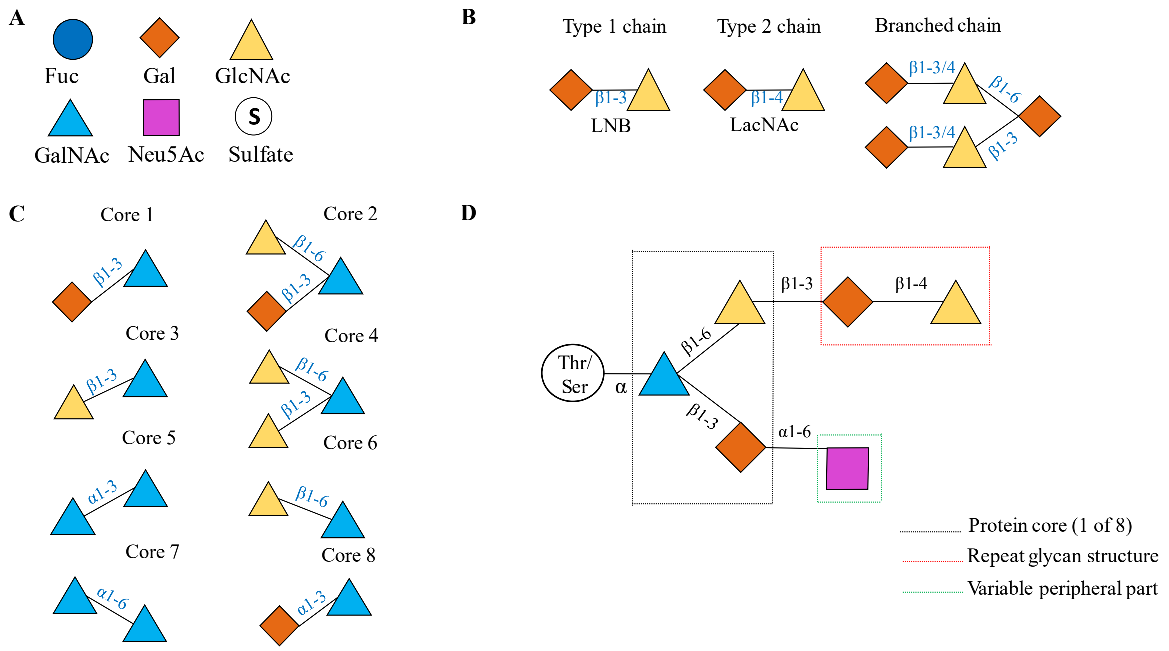

Figure 1.3. Mucin O-glycans.
(A) Monosaccharides and sulfate can be added to the mucin glycans. **(B)** Backbone- repeat glycan structure of mucin. Type 1 chain, type 2 chain & branched chain consisting of Gal and GlcNAc. **(C)** Di- and tri-saccharides in the protein core of mucins and their respective linkages. GalNAc is always bound to Threonine or Serine. **(D)** Example illustration of mucin. The variable peripheral part varies depending on blood group antigens, lewis antigens, and other substituents can be sulfate, fucose or Neu5Ac (sialic acid). Fuc, Fucose; Gal, Galactose; GlcNAc, N-acetylglucosamine; GalNAc, N-acetylgalactosamine; Neu5Ac, N-acetyl neuraminic acid / sialic acid. This figure is inspired by and made based on Li & Chai, (Li & Chai, 2019) and Masi & Stewart (Masi & Stewart, 2022).

1.4.1. Mucin Utilization and Degradation

Mucins are diverse structures as well as HMOs and rely on proteases, sulfatases, and GHs to be degraded (Tailford et al., 2015). As with HMOs, the sialic acids need to be removed from mucins to make them available for other GHs. The removal is performed by sialidases such as the GH33 family, which vary in substrate specificity. The genes involved in the sialic acid metabolism can be found in *nan* gene clusters in several bacteria, making them capable of fully utilizing the sialic acids after they are removed (Tailford et al., 2015). The majority of these live in the mucus regions of the body, which are high in sialic acids, like the lung, bladder, and the gut, and especially the distal colon (Almagro-Moreno & Boyd, 2009). There are variations between species and the genes they encode, and some of them only encode the sialidases, while other have the full set or even lacks the full set of the *nan*-operon (Tailford et al., 2015). GHs necessary for mucin degradation include sialidases of family GH33, the α-L-fucosidases in GH29 and GH95, endo-β-N-acetylglucosaminidases in family GH85, β-galactosidases in GH2, GH42 and GH20, α-N-acetylglucosaminidases in GH89 and the α-N-acetylglucosaminidases in families GH101 and GH129 (Drula et al., 2021; Tailford et al., 2015). Seven of these GH families are also needed for HMO degradation, and the overview of the GH families is illustrated in table 1.1. Because sulfate is present on mucin glycans, sulfatases are also needed in the degradation of mucin. However, the specific sulfatases are poorly investigated and just recently a few sulfatases produced by *Bacteroides thetaiotamicron* were characterized (Luis et al., 2021).

1.5. HMO- and Mucin- Degrading Bacteria

Several bacterial species are known to degrade HMOs and mucins in the human gut. So far, a few species have been studied for their mucin-degradation abilities, such as *Akkermansia muciniphila, Ruminococcus gnavus, Bifidobacterium longum subsp. infantis, Bacteroides fragilis* and *Bacteroides thetaiotamicron* (Tailford et al., 2015). For the utilization of HMO, there seem to be a few common denominators, which are species of the genera *Bacteroides* and *Bifidobacterium* (Masi & Stewart, 2022). Because mucin highly resembles HMO, the knowledge of HMO utilization among other gut commensals that are related to mucin utilization should be emphasized. There is evidence that species of *Bacteroides,* for example, can degrade HMO. For instance, did most of *Bifidobacterium* spp. and *Bacteroides* spp., grow and produce lactate and SCFA when fed certain HMOs in a performed experiment (Yu et al., 2013). Additionally, *B. thetaoitamicron* and *B. fragilis* showed an upregulation of mucin-utilizing genes during HMO consumption, thus indicating HMOs being attractive to more species than those we already know well, such as *Bifidobacterium* species (Marcobal et al., 2011).

Bacteroides

Bacteroides is a gram-negative, anaerobic bacteria, and it is a genus that belongs to the family *Bacteroidaceae* and phylum Bacteroidetes. The GC-content of the *Bacteroides* DNA ranges between 40-48%, and they have a circular genome sized from 2.1 Mb to 7.9 Mb (Wexler, 2014). According to the List of Prokaryotic names with Standing in Nomenclature (LPSN) web interface (Parte et al., 2020), *Bacteroides* is identified as 101 species and eight subspecies, while *Parabacteroides* (formerly *Bacteroides*) is identified as ten species. The Virtual Metabolic

Human Database (VMHD) provides information about the inhabitants of the human gut microbiome; *Bacteroides* are identified as 63 species and *Parabacteroides* as six species in this database today (Magnúsdóttir et al., 2017). Some *Bacteroides* species are known as opportunistic pathogens, acting as beneficial in the right location but a pathogen in other parts of the body and could lead to diseases like oral infections, meningitis, or pericarditis (Zafar & Saier, 2021). One example is the *B. fragilis* which is also highly prevalent in the human gut (Zafar & Saier, 2021). When investigating the gut microbiota of infants, vaginal delivery seems to be a cause for colonization of the *Bacteroides* genus, as vertical transfer from the mother has been suggested by several studies (Backhed et al., 2015; Carrow et al., 2020; Zafar & Saier, 2021). *Bacteroides* members, including *Parabacteroides*, degrade a lot of complex and simple sugars, oligosaccharides, and polysaccharides, including HMOs as well as mucins and plant-derived polysaccharides (Borewicz et al., 2019). The human intestine is abundant in these substrates and is an attractive place for the *Bacteroides* genus to settle, which appear among the most abundant genera in the human gut.

Bifidobacterium

Bifidobacterium is a gram-positive, strictly anaerobic genus that belongs to the *Bifidobacteriaceae* family and Actinobacteria phylum. The GC-content are known to be high with an average of 60%, and the genome size ranges from approximately 1.73 Mb to 3.25 Mb (Milani et al., 2014). The genus was already isolated from infant fecal samples already in the late 1800s (Milani et al., 2016) and is today identified as 106 different species and 18 subspecies in total, according to the LPSN web interface (Parte et al., 2020). According to the VMHD, the *Bifidobacterium* genus is identified as 39 species (Magnúsdóttir et al., 2017).

In the human adult gut, the most common *Bifidobacterium* species are *Bifidobacterium adolescentis* and *Bifidobacterium longum*, but in the gut of infants, the most common *Bifidobacterium* species are *Bifidobacterium bifidum, Bifidobacterium breve* and *Bifidobacterium longum* (Turroni et al., 2012). However, there are no strict boundaries between infant and adult groups as *Bifidobacterium* is thought to be a common vertical transmitted bacteria from mother to child through the mother's vaginal tract or human breast milk (Makino et al., 2013). Since the genus is generally genetically adapted to the utilization of glycans in milk, the milk can also act as a carrier for vertical transmission of bifidobacteria and a good energy source (Milani et al., 2016). This could be used to explain why the genus is not only isolated from the human gut but from the gut of some mammals, birds, and insects whose offspring need parental care as well (Ventura et al., 2014).

Bifidobacterium is not the only one but a well-known and the main commensal HMO-utilizer, and the genus usually accounts for more than 50% of the total bacterial gut population within breast-fed infants (Gotoh et al., 2018). They colonize their host right after birth, and the abundance usually decreases again around six months with weaning and introduction to solid foods and aging. As the enzymes present in the genome of bifidobacterial species degrade complex diet carbohydrates, and host-derived carbohydrates, the products are not only SCFAs such as acetate but also lactate and succinate (Ioannou et al., 2021). These products are beneficial for the growth of other bacteria as they are made available for degradation in their niche and help regulate the dynamics between bacteria in the gut. *Fecalibacterium prausnitzii* utilize acetate through butyrate production as one example of cross-feeding between bifidobacteria and other species (Rios-Covian et al., 2015).

1.6. Short-Chain Fatty Acids (SCFA)

SCFA are known end-products from bacterial fermentation of dietary fibers and polysaccharides that the host cannot digest themselves (Hur & Lee, 2015), and they are important for food intake, inflammations, and insulin signaling. The fermentation of the different SCFAs is performed by enzymes that, like GHs, belong to the CAZyme family.

The main SCFAs produced are acetate, propionate, and butyrate in a 60:20:20 mmol/kg ratio in adults (Martin-Gallausiaux et al., 2021), but depending on the intake of dietary fibers and the presence of carbohydrates, this ratio will differ. Within infants, for instance, the dietary polysaccharides are derived mainly from milk and the microbiota that inhabits the intestine at this time occupy different niches. Branched-chain fatty acids (BCFA) originate from protein and amino acid breakdown (den Besten et al., 2013), and are present at much lower concentrations than the three main fatty acids. These are isovalerate, 2-methyl butyrate, and isobutyrate. Acetate is produced by most gut bacteria, and butyrate and propionate are produced by more specific species in the gut.

For butyrate production, substrates such as acetate, lactate, amino acids, and other carbohydrates are used (Martin-Gallausiaux et al., 2021). For six-carbon sugars, the glycolysis converts monosaccharides into phosphoenolpyruvate (PEP), which is further fermented into alcohols or organic acids. For five-carbon sugars, the pentose-phosphate pathway converts the monosaccharides into PEP (den Besten et al., 2013). Families belonging to the Clostridiales order can produce butyrate, for instance, *Eubacterium, Ruminococcus,* and *Feacalibacterium* (den Besten et al., 2013; Martin-Gallausiaux et al., 2021). As for butyrate, propionate can be produced from lactate and substrates such as 1,2-propanediol, amino acids, and carbohydrates (Martin-Gallausiaux et al., 2021). *Veilonella* and *Bacteroides* can produce propionate using the succinate pathway, while others ferment lactate into propionate through the acrylate pathway (den Besten et al., 2013; Martin-Gallausiaux et al., 2021).

1.7. Analytical Approaches to Study the Taxonomic and Functional Aspects of the Gut Microbiota

The evolution in science has accelerated massively over the past decades, especially considering DNA technologies and bioinformatic tools. Gut microbes are difficult to cultivate as they live under conditions that are challenging to recreate in the lab. From *in vitro* studies and cultivation to understanding different omics – today, extracting DNA and RNA from samples is sufficient to gain knowledge that was not possible before.

Omics is a collective description of biological studies that identify, quantify and investigate the characteristics of genes, proteins, and metabolites in a cell, tissue, or organism – and that ends with -omics (Vailati-Riboni et al., 2017). These are genomics, transcriptomics, proteomics, and metabolomics from genes, mRNA, proteins, and metabolites, respectively. Metagenomics is the collection of genomes that, in theory, exists in a sample (Escobar-Zepeda et al., 2015). Looking at the metagenomes of samples can provide insights into complex ecosystems, as the genes present in each genome play a role here. However, the presence of a gene does not necessarily reflect the activity of the protein. Transcriptional and translational factors such as insertions or deletions may influence the result from gene to protein, deactivating the protein – and it is the study of transcriptomics and mRNA used to characterize the genes and measure how they are regulated (Lowe et al., 2017). Nevertheless, post-translational changes can occur, and therefore, transcriptomic studies might not be sufficient. In this case, proteomics is necessary, which is the study of the function of all expressed proteins (Tyers & Mann, 2003). A perfect example of post-translational changes is reported in the gene *nosZ* that exists in certain soil bacteria and codes for the enzyme N_2OR (Bakken et al., 2012). N_2OR reduces the greenhouse gas N_2O into N_2, but low pH in the soil may result in the protein's deactivation and prevent the reaction from happening. Even if the protein is produced and activated – there is no guarantee that it will function and produce N_2 if the conditions around it are not satisfying. Furthermore, metabolomics is used to measure and compare the metabolites present in a sample, like with the use of gas chromatography (GC) and the analysis of SCFAs.

1.7.1. Bioinformatic Tools

The range for the use of bioinformatic tools is immense, from polymerase chain reaction (PCR) primer design to sequence alignments and protein structure prediction. These tools are used to get higher accuracy, information, and efficiency of the data while also being cost-effective. It is also useful for gene prediction when investigating genetic disorders such as cancer, autism, or diabetes. This is due to the discovery and characterizing of genetic changes in genomes. Accordingly, bioinformatic tools have been used to understand the effects of these changes. Many of the tools can be used to annotate genes and proteins that are present in the amino acid sequences that one can gain from, e.g., shotgun sequencing results. With the help of these tools, it is possible to discover functions, gene names, and other valuable information to gain more insight into the relevant genomes. Examples of these are represented in table 1.2.

Table 1.2. Overview of bioinformatic tools used to annotate genes. This information is withdrawn from the sources listed in the table (date: 15.03.2022).

Tools	**Description**	**Limitations**	**Source**
eggNOG mapper v2 (Batch Functional Annotation)	Uses precomputed Orthologous Groups (OGs) and phylogenies from the EggNOG database to transfer functional information from fine-grained orthologs only.	Up to 100 000 proteins in FASTA format	(Cantalapiedra et al., 2021) (Huerta-Cepas et al., 2018)
dbCAN	Annotate proteins (FASTA formate) using DIAMOND, HMMER, and eCAMI via CAZy, dbCAN, and CAZyme peptides, respectively.	Max 20MB file size	(Zhang et al., 2018)
GhostKOALA & BlastKOALA	Complete KO (KEGG orthology) assignments to characterize individual gene functions. Reconstruct KEGG pathways, BRITE hierarchies, and KEGG modules to imply high-level functions of the organism or the ecosystem.	1-30 proteins with max. length of 40 000 aa	(Kanehisa et al., 2016)
InterProScan	Software package that functionally characterizes nucleotide or protein sequences.	1-30 sequences	https://www.ebi.ac.uk/interpro/about/interproscan/

1.7.2. Analyzing Proteins

Extraction of proteins can be done from various samples like the soil, humans, plants, animals, viruses, etcetera. The cells must be lysed to separate the protein from the host cell. This process is easier in cells of mammalian cells, as the plasma membrane is easily disrupted, while the cell wall appears more rigid in fungi and bacteria (Tan & Yiap, 2009). When performing a mass spectrometry (MS)-based proteome analysis, the proteins must be converted to peptides first (Wiśniewski et al., 2009). This process involves a solubilization of the protein with detergents, separation by sodium dodecyl sulfate (SDS) polyacrylamide gel electrophoresis (PAGE), and digestion of the trapped proteins (Wiśniewski et al., 2009). Proteins treated with SDS will unfold and the SDS molecules binds to the protein almost proportional to the protein mass. Accordingly, when running the proteins through the gel in PAGE they will be separated based on their mass as SDS is negatively charged. This causes the protein to have a charge density and it will be driven through the gel with the same force, making small molecules migrate faster than bigger molecules (Tan & Yiap, 2009). The PAGE is not only used to separate specific proteins, but also to purify them by removing contaminants. The advantage of having the proteins in a gel is the reduced risk of contaminations (Wiśniewski et al., 2009). After SDS-PAGE, the proteins can be visualized in the gel by staining with Coomassie Blue. The wanted proteins can then be extracted from the gel for further analysis.

Mass spectrometry (MS) is the most used method to identify the proteins in a sample, and it can also be used to look at post-translational modifications in proteins, like methylation. With the extensive analysis of samples and high throughput, genomics provides complete genomic sequences, which are important to identify proteins quickly and correlate to MS measurements of peptides (Domon & Aebersold, 2006). There are different mass spectrometers, and a well-known method is the LC-MS/MS (liquid chromatography-tandem mass spectrometry). Briefly explained, the spectrometer has an ion source, a mass analyzer, and a detector to detect the proteins ion mass. In LC-MS/MS, the proteins are converted into a gaseous state, and by measuring the mass to charge ratio, they are characterized by high resolution and mass accuracy. The result is a mass spectrum that can be used to identify the peptides.

1.7.3. Sequencing Methods for Nucleic Acids

Different approaches to investigate the gut microbiota composition and its functions are limited when considering the different sequencing methods. In theory, all genomes in a sample would be represented using the available technologies today. However, 16S rRNA gene sequencing can provide poor taxonomic resolution challenging the accuracy of similar taxonomies (Nilsen et al., 2021; Ravi et al., 2018), while Shotgun sequencing is a more accurate method but expensive and complex due to storage and analysis complications. Several technologies have been developed since the mid-seventies, and DNA sequencing methods have evolved rapidly.

Sanger sequencing is known as a first-generation sequencing method, and it was developed back in 1977 by Frederick Sanger and his colleagues – and is the first method in DNA sequencing (Heather & Chain, 2016). It is based on dideoxynucleotides (ddNTP) in DNA polymerization reactions, and the ddNTP has one less OH-group than the deoxynucleotide (dNTP) in DNA. dNTP is the substrate for polymerase, and if the 3'OH group on the nucleotide is removed, a new nucleotide will not attach, and the reaction will stop (Heather & Chain, 2016). The template for sequencing is divided into four tubes, and each tube has a primer, dNTPs, DNA polymerase, and one radiolabeled ddNTP. This provides a normal polymerization reaction in each tube until the ddNTP attaches and terminates it. Knowing the length of the fragment where termination occurred, the position of the added ddNTP is known. This is visualized with polyacrylamide gel electrophoresis and radiography, and the fragments can be read from the gel picture in the correct order. Today, each ddNTP has a fluorescence dye, making it possible to use only one tube for the reaction and capillary-based electrophoresis. When high throughput is not necessary, Sanger sequencing is sufficient and is good for specific primers on specific templates, such as plasmids or PCR products, as it read lengths up to 1000bp (Heather & Chain, 2016; Slatko et al., 2018).

Next-generation sequencing (NGS), or second-generation sequencing, has several platforms, such as Illumina, 454 (Pyrosequencing), and Ion Torrent (Slatko et al., 2018). The platforms have small method variations, but all apply to the same concept. The method differs from Sanger sequencing as it does not apply radiolabelling or fluorescence dye before visualization on a gel (Slatko et al., 2018). Instead, samples are prepared in a library with amplified DNA or ligation with custom adapters. The library is then applied to a solid surface, amplifying each fragment with covalent links that hybridize the library adapters. This results in clusters of DNA, each coming from a single DNA fragment that will be a single sequencing reaction. In Illumina, the nucleotides are marked with a fluorophore that will emit a signal when it is cut off, revealing the correct sequence order (Illumina, 2017). A big advantage of using NGS is the need for a single volume on the reaction plate where all reactions are run in parallel, making it a very cost-effective method. However, they rely on shorter reads up to ~ 500 bp (Heather & Chain, 2016).

There are also third-generation sequencing methods which still is relatively new. It aims to sequence long DNA and RNA molecules and is therefore also known as Large Fragment Single

Molecule (LFSM) sequencing or Single Molecule Sequencing (SMS) (Heather & Chain, 2016; Slatko et al., 2018). Examples of platforms using these methods are PacBio and Oxford Nanopore, which can sequence full genomes in a short amount of time. The backsides of third-gen technologies are their relatively high error rate and incredibly vast storage needs.

1.7.4. Analyzing the Short-Chain Fatty Acid Composition

The SCFAs in the human intestine are not produced by human cells as humans lack the enzymes needed to degrade the fiber substrates (den Besten et al., 2013). Instead, they are produced by the microbial community in the intestine, providing the necessary fuel for intestinal epithelial cells. SCFAs also regulates epithelial cell functions and strengthen the gut barrier functions (Martin-Gallausiaux et al., 2021). The colonocytes absorb most of the SCFA in the cecum and large intestine. This leads to an issue when analyzing the SCFA composition in the intestine because only 5% are secreted in the analyzed feces (den Besten et al., 2013). Methods to analyze and study the SCFA composition in feces have advanced over the decade, and gas chromatography (GC) is the most common and precise method used (Primec et al., 2017). Other known methods are related to liquid chromatography (LC), like high-performance liquid chromatography (HPLC), nuclear magnetic resonance (NMR), and capillary electrophoresis (CE) (Primec et al., 2017).

Gas chromatography is compatible with SCFAs due to the acid’s volatile properties. The GC consists of a stationary phase and a mobile phase, and is coupled to a detector, which collects the data analyzed by a computer. In the mobile phase, the carrier gas that the samples are separated by, interacts with the stationary phase (Primec et al., 2017). The carrier gas varies based on the column and the detector used, but typically they are helium, argon, hydrogen, or nitrogen. The samples are loaded into a column where the stationary phase is normally based on polysiloxanes or polyethylene glycol (PEG) (Primec et al., 2017). The most used detector is the flame ionization detector (FID), sensitive to ionized hydrogen molecules. The ions are creating a current proportional to the organic compounds in the sample, and this current is registered by the detector and is then analyzed by a computer program creating chromatograms.

1.8. PreventADALL

Different studies have investigated the bacteria to host interactions for multiple years, intending to prevent diseases in humans and improve public health. One of them is the Preventing Atopic Dermatitis and ALLergies in children (PreventADALL) – study. In this study, the goal has been to prevent allergic disease development in early infancy, as atopic dermatitis for instance, that may predispose to food and other allergy development later in life (Lodrup Carlsen et al., 2018). Early life factors, exposure, environment, microbiota, and xenobiotics have been assessed in the study and the biological sampling included blood, urine, skin swabs and feces for microbiota, placental biopsies and swabs, amniotic fluid (if CS), vernix caseosa, saliva and breast milk (Lodrup Carlsen et al., 2018). A total of 2397 mother-child pairs from Oslo, Østfold and Stockholm were enrolled in the study.

1.9. Aim of Thesis

Bacterial utilization of HMO- and mucin-related substrates have been thoroughly investigated in vitro. However, there is a knowledge gap related to the metabolism and potential competition for these resources between *Bifidobacterium* and *Bacteroides* in the infant gut. Both genera have been observed to utilize both mucin and HMO, but their metabolism and potential benefits remain unclear.

This thesis aims to investigate the metabolic pathways and glycoside hydrolases known for the degradation of mucins and HMOs in *Bifidobacterium* and *Bacteroides*. To achieve this, the following sub aims were included:

- Create a database containing the genes only from *Bifidobacterium* and *Bacteroides* species and compare the presence of these genes between the two genera
- Identify proteins involved in HMO- and mucin-utilization with proteome analysis
- Examine short-chain fatty acid composition from gas chromatography

This will be addressed by analyzing the genome and proteome data gathered from *Bifidobacterium* and *Bacteroides*, using the feces of 6-month-old infants collected from the PreventADALL study.

2. Materials and Methods

An overview of the performed experiments is illustrated in figure 2.1.

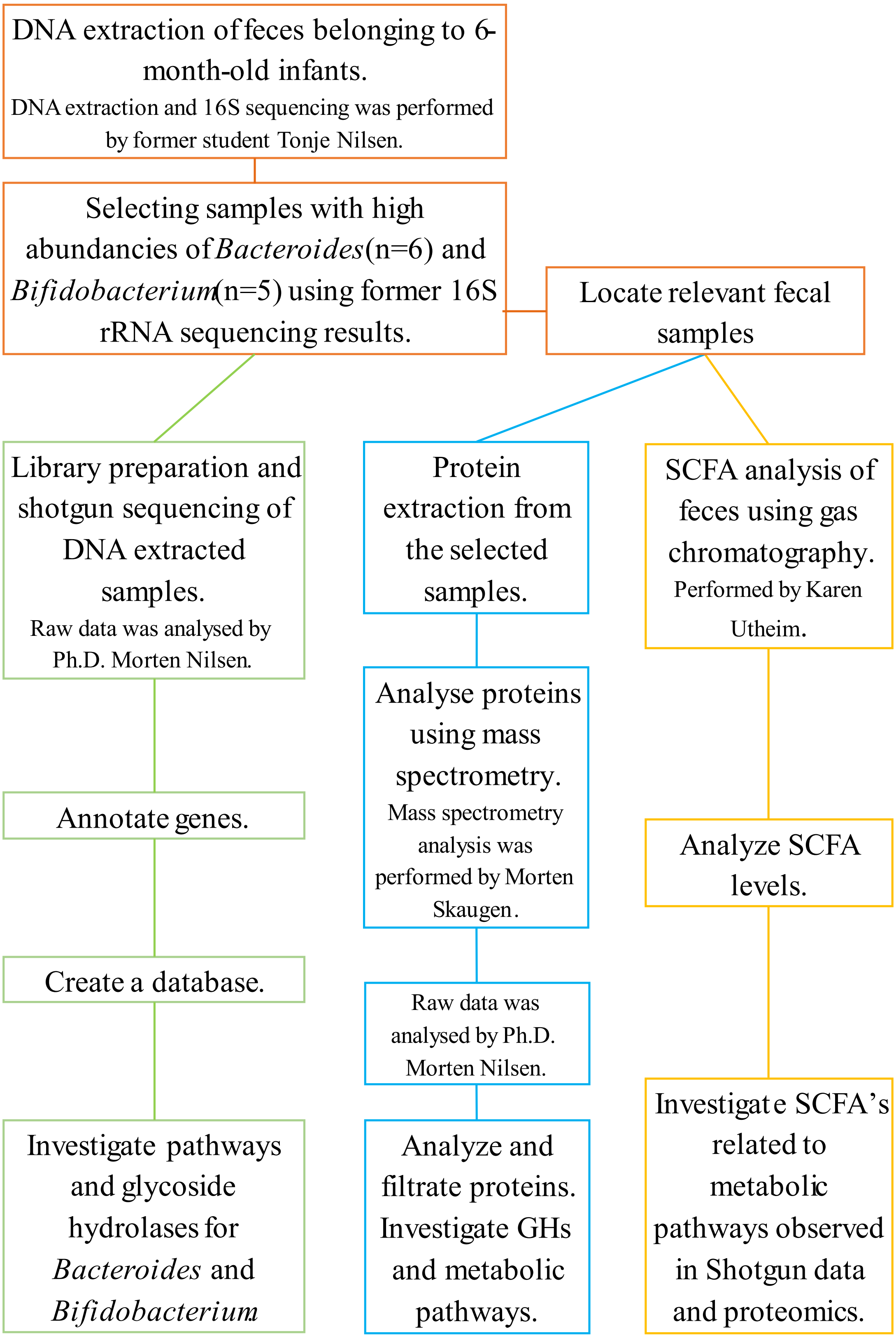
DNA extraction of feces belonging to 6-month-old infants.
DNA extraction and 16S sequencing was performed by former student Tonje Nilsen.
Selecting samples with high abundancies of *Bacteroides* (n=6) and *Bifidobacterium* (n=5) using former 16S rRNA sequencing results.
Locate relevant fecal samples
Library preparation and shotgun sequencing of DNA extracted samples.
Raw data was analysed by Ph.D. Morten Nilsen.
Protein extraction from the selected samples.
SCFA analysis of feces using gas chromatography.
Performed by Karen Utheim.
Analyse proteins using mass spectrometry.
Mass spectrometry analysis was performed by Morten Skaugen.
Annotate genes.
Analyze SCFA levels.
Create a database.
Raw data was analysed by Ph.D. Morten Nilsen.
Investigate pathways and glycoside hydrolases for *Bacteroides* and *Bifidobacterium*
Analyze and filtrate proteins. Investigate GHs and metabolic pathways.
Investigate SCFA's related to metabolic pathways observed in Shotgun data and proteomics.

Figure 2.1. Flow chart of the experiments. The flow chart illustrates the workflow in this thesis.

2.1. Sample Criteria

The following criteria were set when selecting the samples used in the experiments: There had to be enough feces left in the samples for both the proteome analysis and short-chain fatty acid analysis. The fecal samples had to have a relatively high abundance of either the *Bacteroides* or *Bifidobacterium* genus. All samples were from 6-month-old infants that were breastfed until at least six months of age, and all infants were delivered vaginally. A total of 14 samples were selected for shotgun sequencing, six with high amounts of *Bacteroides* and five with high amounts of *Bifidobacterium*, including one positive (*Bifidobacterium Breve)* and two negative controls (negative from DNA extraction plates + PCR-water). The abundance of *Bacteroides* and *Bifidobacterium* in these samples can be found in Appendix A1, table 1.

2.2. Short-Chain Fatty Acid Analysis of Samples with Gas Chromatography

Fecal samples were 10 × diluted in $_dH_2O$, whereas 300 µl of the sample was diluted 1:1 with an internal standard (0.4 % formic acid and 2000 µM 2-methyl valeric acid). The samples were centrifuged at 13 000 rpm for 10 minutes, and the supernatant was filtered through spin columns (0.2 µM filter, VWR USA), and centrifuged at 10 000 rpm for 5 minutes. The eluate was transferred into a GC vial (VWR, USA). Trace 1310 with autosampler (ThermoFisher Scientific) was the gas chromatographic instrument used. The data program used to identify the peaks was Thermo Scientific^{TM} Dionex^{TM} Chromeleon^{TM} 7 Chromatography Data System Version 7.2 SR4. For more details, see Appendix B, protocol B1.

2.3. Extraction of DNA from Fecal Samples

The following procedures for DNA extraction from fecal samples had already been performed by lab personnel using the “MagPure Stool DNA LQ kit” (Angen Biotech, China). The kit was used according to the manufacturer’s recommendations.

Before starting, the extraction kit contents had to be prepared. This was done by adding PDB to the Proteinase K to a 20 mg/ml final concentration, which was then stored at -20 to +8˚C. PDB was then added to RNase A to a 15 mg/ml final concentration, also stored at -20 to +8˚C. Buffer GW1 was diluted with 100 % ethanol and stored at room temperature. Buffer MLE was diluted with isopropanol and stored at room temperature. Buffer ATL had PVP-10 powder added to it before use, and Magnetic Particles N was shaken properly before use.

To isolate the bacterial DNA, lysis was performed by transferring 100-150 mg sample (150 ml if liquid) to a 2 ml bead tube. Buffer ATL+PVP-10 and Buffer PCI were added to the samples.

Samples were run on FastPrep96 (MP Biomedicals) for 2 × 40 seconds at 1800 rpm with a 5-minute break between rounds. The samples were incubated on a heat block at 65˚C for 20 minutes and should be fully lysed. Centrifuging at 13 000 × g for 5 minutes was performed to collect bigger particles. The samples were then stored at 4˚C until further procedures.

The sample was transferred to a KingFisher Deep Well (DW) plate. The samples were treated with RNase A for 10 minutes at room temperature, to remove potential RNA contaminations. MagPure Particles N is paramagnetic beads that bind the negatively charged DNA. Proteinase K denatures proteins, and lysis buffer ensures viscosity and the correct pH. The KingFisher plates were prepared manually with the recommended volumes. Proteinase K, Buffer MLE, MagPure Particles N, and sample + RNase A were mixed. The following steps were performed by the KingFisherFlex robot (Thermo Fisher Scientific, USA) to extract the DNA. The first sample wash was performed using Buffer GW1, and the two next wash rounds were performed using ethanol. The MagPure particles released DNA in the last step by adding elution buffer to the sample.

Illumina 16S sequencing was performed on all samples, and the data was used to investigate the abundance of *Bacteroides* and *Bifidobacterium* in all samples, as explained in the sample criteria (section 2.1). The extracted DNA was left at -20˚C and was used to prepare for Shotgun sequencing.

Before continuing with the samples, it was important to ensure that DNA was present in the selected samples. The presence was checked by measuring the DNA concentration and checking the samples on a gel.

2.4. Measuring DNA Concentrations using Qubit

The DNA concentration was measured using the Qubit Quant-iT^{TM} Assay (Thermo Fisher Scientific, USA) to ensure sufficient DNA in the following library preparations. The Qubit^{TM} fluorometer (Qubit 9V, Invitrogen, USA) was used to quantify DNA before tagmentation, after library clean-up, and after pooling of libraries. 198 µl of Working Solution (1:200 dilution of Quant-iT^{TM} reagent in Quant-iT^{TM} buffer) was added to 2 µl of DNA sample. All samples were then vortexed and incubated in the dark at room temperature for approximately 2 minutes before being measured in the Qubit^{TM} fluorometer.

2.5. Agarose Gel Electrophoresis

Gel electrophoresis was used for quality assessment and the presence of the DNA in the samples. The gel was a 2 % agarose gel (with peqGreen), which was run on 80 V for 42 minutes with a 100 bp ladder after the samples were amplified. When running the gel later after library pooling of shotgun samples, it was run at 80 V for 50 minutes with the same sized ladder. The gels were visualized using the Molecular Imager Gel DOC^{TM} XR Imaging Systems (appendix A2, Figure 1).

2.6. Preparing Samples for Shotgun Sequencing

According to the manufacturer's recommendations, the following steps were performed using the ‘Illumina® DNA Prep, (M) Tagmentation’ protocol.

Tagmentation of genomic DNA was performed using Bead-Linked Transposomes (BLT), which fragment and tag the DNA with adapter sequences. Tagmentation Buffer 1 (TB1) ensured the correct pH. 54.4 ng to 438 ng DNA was mixed with the tagmentation master mix, including BLT and TB1. The tagmentation reaction was induced by running the following program on the thermal cycler: Preheat lid option set to 103˚C, reaction volume set to 50 µl, 55˚C for 15 minutes, and then hold at 10˚C.

Before the samples were ready for amplification, the tagmentation reaction had to be stopped by adding Tagment Stop Buffer (TSB). TSB may precipitate but is easily dissolved by heating at 37˚C for approximately 10 minutes. The samples were then sealed and run on the following program on the thermal cycler: Preheat lid option set to 103˚C, reaction volume set to 60 µl, 37˚C for 15 minutes, and hold at 10˚C. When placed on the magnetic stand for approximately 3 minutes, the supernatant was removed, and the beads were washed twice with 100 µl Tagment Wash Buffer (TWB). Then TWB was added again to cover the beads to prevent them from drying out.

TWB supernatant was removed, and PCR master mix was added. The PCR Master Mix included enhanced PCR mix (EPM) and PCR-water. The samples were mixed with a pipette to resuspend the beads fully, and droplets were collected with a quick spin. Index adapters were then added to each sample. The tagmented DNA was amplified using a combination of four i5 and four i7 adapters (Appendix A1, Table 2). The samples were sealed and centrifuged, and placed on the thermal cycler with the following PCR amplification program: Preheat lid option was set to 103˚C, then 68˚C for 3 minutes, and 98˚C for additionally 3 minutes. They were then run on five cycles of 98˚C for 45 seconds, 62˚C for 30 seconds, and 68˚C for 2 minutes. Finally, the temperature was set to 68˚C for 1 minute and then held at 10˚C. 5 µl of each sample was collected for gel electrophoresis, and the rest was left at 2-8˚C overnight.

Each sample library was purified after amplification using Sample Purification Beads (SPB), Resuspension Buffer (RSB), and 80% EtOH. The samples were centrifuged and placed on the magnetic stand for 5 minutes, and 40 µl supernatant was transferred to a new plate. A clean-up with a 1.8 × SPB– to– sample ratio was performed, and the beads were washed twice with 200 µl EtOH before RSB was added to eluate the DNA, which was transferred to a new plate. An exception was two of the samples where RSB was added initially instead of SPB, whereas the ratio of SPB was adjusted to fit the new volume. The samples were stored overnight at -25 to -15˚C.

Before pooling the libraries, the samples were normalized by measuring Qubit concentrations. The amount of sample added to the library ranged from 2 µl for the highly concentrated samples to 21 µl for the less concentrated samples, to add 80 ng of DNA from each sample. The final library concentration measured from Qubit was approximately 9,82 ng/µl.

Finally, the pooled library was checked on an agarose gel as previously described and sent to NovoGene (UK) for Illumina sequencing. The gel picture can be found in Appendix A2, figure 2.

2.7. The Proteome Analysis

Isolating bacterial cells

The same samples that were used for shotgun sequencing, were used for the proteome analysis. Approximately 0.2 grams of feces were weighed out for each sample, including replicates (Appendix A, Table 3). Fecal samples were suspended in a cold TBS-buffer and passed through a Merck^{TM} Nylon-Net Steriflip^{TM} Vacuum Filter Unit (20 µm, Fisher Scientific, USA) to remove fibrous materials and human cells. Samples were centrifuged at 4000 rpm for 10 minutes to pellet bacterial cells, and then resuspended in TBS. The second filtering to remove eukaryote proteins and capture bacterial cells was performed using the Millipore Vacuum unit (Merck Millipore, USA) and 0.22 µL membrane filters (Millipore, USA).

Lysis of bacterial cells

Fast-Prep tubes were prepped with Lysis buffer (50 mM Tris HCl, 200 mM NaCl, 0.1% Triton-X100, 10 mM Dithiothreitol, and 2% Sodium Dodecyl Sulfate) and acid-washed glass beads

(0.2 g of <106 µm beads, 0.2 g of 425-600 µm beads, and 2 × of 2.5-3.5 mm beads (Sigma-Aldrich, Germany)) together with the membrane filters containing the sample. A negative control was also included containing only lysis buffer and beads. All tubes were put on ice for 30 minutes and occasionally vortexed. Cells were disrupted using FastPrep96^{TM} (MP Biomedicals, USA) by 3 × 60-second pulses at 1800 rpm and then centrifuged at 16 000 g for 15 minutes at 4˚C. The supernatant was then transferred to new tubes (approximately 700 µl).

Bicinchonic Acid (BCA) protein assay - measuring protein concentration

All samples, including the blank (lysis buffer), were diluted at 1:5 in $_dH_2O$. BCA working solution from the Pierce BCA Protein Assay Kit (ThermoFisher Scientific, USA), which had a reagent-to BCA-ratio of 1:50, was added to the samples. The samples were then incubated at 60˚C for 30 minutes before being transferred on ice to stop the reaction. The absorbance was measured at 562 nm on the Eppendorf BioPhotometer D30 (Eppendorf AG, Germany), which was calibrated with BCA standard solutions (25, 50, 100, 150, 200, and 250 µg/ml). All sample measurements can be found in Appendix A1, Table 4.

SDS-PAGE – purification of protein

All samples were up-concentrated based on the previous BCA measurements using speedvac to the desired volume of 19.5 µL containing approximately 40 µg protein for the best results on the mass spectrometer (MS). This is equivalent to 2.05 µg/µL protein, which was used on the SDS-PAGE (Sodium Dodecyl Sulfate – Polyacrylamide Gel Electrophoresis). SDS-PAGE was used as a clean-up step to purify the proteins and get rid of other contaminants.

A heating block was preheated to 90˚C, and a reducing sample buffer was made, using 4 × sampling buffer (ThermoFisher Scientific, USA) and 10 × reducing agent (ThermoFisher Scientific, USA). The sampling buffer binds to the proteins and colors them in the gel later. The reducing sample buffer was added to the 40 µg protein sample and was incubated at 90˚C for 5 minutes. DTT in the reducing agent keeps the protein unfolded in this denaturing step. All samples were centrifuged at 10 000 g for 1 minute. The gel (Mini-PROTEAN TGX Stain-Free Gels, Bio-Rad Laboratories, USA) was unpacked and assembled. Freshly made 1 × TGS buffer (Tris-Glycine-SDS, Bio-Rad, USA) was poured into the inner chamber while used 1 × TGS buffer was poured into the outer chamber. 30 µl of each sample was transferred to wells in the gel using every other well to avoid contamination of samples. The gel was run at 270 V for 5 minutes. Gel 1 was removed after 5 minutes, and gels 2 and 3 were run for additional 4 minutes to separate all the samples from the well. Also, gels 4 and 5 were run for three additional minutes.

Staining and de-staining of SDS-gel

A 1000 mL stock with a de-staining solution was made containing 25 % isopropanol, 10 % glacial acetic acid, and Milli-Q-water. A 0.05 % Coomassie staining solution was made (Coomassie Brilliant Blue R250, Bio-Rad, USA) in a 100 mL destaining stock. All gels were stained for 1 hour at 20 rpm, then de-stained for 2 × 20 minutes at 20 rpm in destaining solution, and were then de-stained overnight at 20 rpm in a 1:2 destaining stock solution. The colored bands of the gel samples were cut into 1 × 1 mm cubes, 200 µl $_dH_2O$ were added to cover the gel pieces and the samples were stored in the refrigerator for a couple of days.

In-gel reduction, alkylation, and digestion

The samples were incubated for 15 minutes at room temperature at 500 rpm on a Thermo mixer. The liquid was removed and 200 µl 50% ACN (Acetonitrile, Sigma-Aldrich) / 25 mM AmBic (Ammonium Bicarbonate, Sigma-Aldrich, USA) was added. The samples were incubated for 15 minutes at room temperature and 500 rpm. The liquid was removed and 200 µl $_dH_2O$ was added again before the previous steps were repeated. Afterward, 100 µl of 100 % ACN was added, and samples were then incubated for 5 minutes at room temperature at 500 rpm. The liquid was removed, and lids remained open to air dry samples for approximately 2 minutes.

The samples were reduced, meaning the disulfide bonds were cleaved, by adding DTT solution (10 mM DTT (Dithiothreitol, Sigma-Aldrich, USA), 100 mM AmBic) and incubated for 30 minutes at 56˚C at 500 rpm. The samples were cooled down and DTT solution was removed before IAA solution (55 mM IAA (Iodoacetamide, Sigma-Aldrich, USA), 100 mM AmBic) was added, and the samples were incubated for 30 minutes in the dark at room temperature. IAA prevented the proteins from forming disulfide bonds. IAA solution was removed and 100 % ACN was added to dry out the gel pieces. Following, the samples were incubated for 5 minutes at room temperature at 500 rpm, and the liquid was removed from the samples which were air-dried as previously described.

Trypsin buffer (25 mM AmBic, 10 % ACN and milliQ-water) was made. For digestion, 30 µl of 10 ng/µl Trypsin solution (0.5 ng/µl Trypsin, Trypsin buffer) was added to each sample which was then incubated for 30 minutes on ice. Trypsin is a Serine protease that cleaves the protein at the carboxyterminal of Arginine and Lysine residues, which is advantageous for MS analysis as it results in a positive charge at the C-terminus of the peptide (Dau et al., 2020).

Additional trypsin buffer was added to the samples to cover the gel pieces, and they were then incubated overnight at 37˚C and 500 rpm. The next day, samples were cooled down, and a 1 % TFA (Trifluoroacetic acid, VWR, USA) was added to terminate the reaction. The samples were then stored in the refrigerator.

Zip Tip and elution of proteins

The samples were centrifuged at 21 500 rpm for 3 minutes to collect droplets. They were sonicated in a water bath for 15 minutes to extract the proteins from the gel and into the TFA solution. New Eppendorf tubes were prepared with 70 % ACN/0.1 % TFA to elute the proteins. The extraction was done using a C_{18} solid-phase extraction method. To enhance the binding of the proteins to the ZipTip (Merck-Millipore, USA), the C_{18} material inside must be equilibrated and conditioned. This was done by pipetting and discarding 100 % MeOH, a 70 % ACN/0.1 % TFA solution, and 0.1 % TFA, whereas the latter is an ion-pairing agent. The sample was then pipetted up and down to bind to the hydrophobic C_{18} material. 0.1 % TFA was pipetted and discarded to cleanse the proteins. They were eluted in the new Eppendorf tubes by pipetting up and down in the 70 % ACN/0.1 % TFA solution. The samples were then dried out with a speedvac and left in the refrigerator.

Mass spectrometry analysis of proteins

Before the mass spectrometry (MS) analysis was performed, peptides were dissolved in 2 % ACN/0.1 % TFA, and 1.5 µl of each sample was measured on a Thermo Scientific NanoDrop One Microvolume UV-Vis Spectrophotometer (A205) (ThermoFisher, USA) using a drop of 2 % ACN/0.1 % TFA as blank. The measurement results are in Appendix A1, table 5. The peptide samples were analyzed by coupling a nano UPLC (nanoElute, Bruker) to a trapped ion mobility spectrometry/quadrupole time of flight mass spectrometer (timsTOF Pro, Bruker). The peptides were separated by an Aurora C18 reverse-phase (1.6 µm, 120 Å) 25 cm × 75 µm analytical column with an integrated emitter (IonOpticks, Melbourne, Australia). Complementary information can be found in Appendix B1, protocol 2.

2.8. Bioinformatic Analysis

2.8.1. The Shotgun Database

Analysis of raw data, annotation of bacterial taxonomy, and creation of FASTA file

For the raw data analysis, the shotgun pipeline that was used, used trimmomatic for adapter clipping and quality trimming. BowTie2 with the human genome was used to map away all human DNA sequences, to ensure the anonymity of all individuals in this study. SPAdes-meta was used to assemble the reads to contigs, and MaxBin and MetaBat2 were used to bin the contigs to metagenome-assembled genomes (MAGs). Drep dereplicates the bins from Maxbin and MetaBat2, and the best candidates will be used further in the analysis. Prodigal was used to convert the DNA sequences to amino acid sequences, which will be mapped to proteins by eggNOG mapper (v2, Batch Functional Annotation). FASTQ was used to determine the quality of the bin, and FASTQC quality-checked the reads, bins, and MAGs. From Drep, bacterial taxonomy was annotated to the contigs within the MAGs using the Kraken algorithm with the HumGut database. R studio version 1.4.1103 was used to annotate the taxonomy of all nodes present in the Shotgun data (Appendix C1). All nodes from *Bacteroides* and *Bifidobacterium* were extracted (Appendix C2), and a FASTA file with the amino acid sequences for the applicable nodes was made (Appendix C3).

EggNOG mapper – to annotate potential proteins from the shotgun database

EggNOG mapper version 2.1.6 annotated all potential proteins present from the FASTA files for *Bacteroides* and *Bifidobacterium* (Cantalapiedra et al., 2021). EggNOG mapper provided information about e-values, taxonomy, EC-numbers, description of proteins, COG-category, gene names, and pathways they belonged to in a table. Nodes that had e-values > 1e^{-10} were removed from the table. The taxonomy from the shotgun data and HumGut database was used for annotation instead of the taxonomy from eggNOG mapper. The code is attached as an R markdown file in Appendix C4.

KEGG Mapper Reconstruct – to reconstruct metabolic pathways

KEGG Mapper – Reconstruct (updated 01.07.2021) was used to reconstruct KEGG pathways with a set of K numbers extracted from the table retrieved from eggNOG and processed in Rstudio (Kanehisa & Goto, 2000). This way, potential proteins were visualized in pathways for the two bacterial genera separately, making it easier to compare them. The code can be found in Appendix C5 as an R markdown file.

dbCAN – to annotate glycoside hydrolases to potential proteins from the shotgun database

All amino acid sequences from the FASTA file retrieved in the beginning, were run through dbCAN to annotate GH families. Tools used in dbCAN were HMMER, DIAMOND, and eCAMI. A threshold value was set for the GH-families having to be equal in at least two of the three tools before annotating them to the node. The rest were removed from the dataset. These annotations were then added to the eggNOG-table and used to identify nodes with GH-families related to HMO- and mucin- degradation. The code can be found in Appendix C6 as an R-markdown file. Afterward, all GHs related to HMO and mucin degradation were counted for each genus.

Sequence alignments of GalT1 and GalT2 in ClustalW

GalT exists as two proteins: GalT1 and GalT2 but is only identified as UDP-glucose-hexose-1-phosphate uridylyltransferase and annotates to the same EC-code (EC 2.7.7.12). The information provided by the online databases was not detailed enough to identify GalT1 and GalT2, but we wanted to identify each of these in the shotgun data as they typically appear in different pathways. An identification of the different GalT protein sequences was performed by Turroni et al. (De Bruyn et al., 2013; Turroni et al., 2010), where the sequence used of *Bifidobacterium bifidum* was deposited in the GenBank database. By doing a sequence alignment in ClustalW with the GalT1- and GalT2-sequences found in the Turroni-article, and two sequences of GalT (EC 2.7.7.12) from *Bifidobacterium bifidum* JCM 1255 in the shotgun database, the GalT proteins could be identified.

Metagenome Assembled Genomes Taxonomy // Data treatment

Only the bins and nodes with annotated species from the *Bacteroides* and *Bifidobacterium* genera were used, and *Phocaeicola* and *Parabacteroides* were included. The database was used to annotate the taxonomy to the MAGs, and MAGs with no nodes from the mentioned groups were removed. At the species level, a threshold value was set for MAGs, which were considered “pure” when one specie represented 80% or more of the sample. This was checked by looking at the table “Table_nodebin” made in Appendix C2, which had information on all nodes, and which sample they came from. Length and coverage of the nodes were also considered. The MAGs that were not considered pure were grouped as “*Bacteroides*,” “*Bifidobacterium*,” or “uncertain” instead of at species level. Samples that had lost significant amounts of their explained abundances were removed (4 and 5 removed) when calculating the average abundance of species in samples high in *Bacteroides* (1-6) and samples high in *Bifidobacterium* (7-11).

2.8.2. Proteome-Analysis after Mass Spectrometry

The raw files from mass spectrometry were analyzed using MaxQuant version 1.6.17.0, and the MAXLFQ algorithm was implemented for label-free quantitative detection of proteins. The sequence database made in Rstudio and the human genome (*Homo Sapiens*, 73 952 sequences) were used to search against the raw files, to get hits on proteins that could derive from *Bacteroides* and *Bifidobacterium*, and to reduce contaminants, respectively.

Filtering of protein data using Perseus v1.6.6.0

Perseus version 1.6.6.0 was used to process further the data that was retrieved from MaxQuant. Rows were filtered based on categorical columns to remove proteins only identified by site, reverse proteins, and potential contaminants. Rows were then filtered based on text columns to remove all human proteins from the human genome (*Homo Sapiens,* 73 952 sequences) to reduce further contaminants. The values were then log2(x) transformed and categorical annotations for the samples were added. One of the categorical annotations described the replicates, and the other was split into: samples high in *Bacteroides* (1-6) and high in *Bifidobacterium* (7-11). Rows were then filtered based on valid values, removing all proteins that appeared once in the dataset. Missing values NaN were replaced by 0 for improved functionality of the dataset.

The shotgun database with information from eggNOG, KEGG, and dbCAN was annotated against nodes in the protein dataset, to match the potential proteins to proteins detected in the proteome analysis. The average value of all replicates was calculated and used instead of each replicate separately. For clustering, all samples that had a low total amount of label-free quantification (LFQ) intensity were removed (<1000) to avoid clustering based on the number of proteins. The LFQ-intensity aims to determine the amount of proteins in a sample. The Pearson correlation was also investigated in scatter plots of each replicate.

KEGG mapper reconstruct – to reconstruct metabolic pathways

KEGG Mapper – Reconstruct was used to investigate pathways found in the proteomics data by extracting the K numbers from the table in Perseus. The lists with K numbers were made for *Bacteroides* and *Bifidobacterium* separately to compare them to the reconstructed pathways made for potential proteins, as described for the shotgun data in Appendix C5. GH-families annotated from dbCAN in Perseus were also compared to the potential presence of proteins in the database.

2.9. Statistical Analysis

Correlation of SCFAs and 16S taxonomy

A matrix with absolute values of SCFA concentration in mmol/kg feces and relative abundances (0-1) of taxonomies from the 16S data was made. To correlate SCFAs to bacterial taxa gathered from 16S data for samples 1-11, spearman's correlation was used, with the significance level set to 0.1 ($p < 0.1$). Through Spearman's rank correlation coefficient rho (ρ), a new matrix was created. P-values were also included as a separate matrix, and together these made a correlation plot. Two additional correlation plots were made but investigated the groups separately (high in *Bacteroides* and high in *Bifidobacterium*). The same method as previously described was used but with a p-value < 0.05. The code for the analysis of the first correlation plot can be found in appendix C7 as an R Markdown file.

3. Results

3.1. Sample Selection

The 16S sequencing results were investigated based on the sample criteria described in section 2.1. Out of 100 fecal samples, 48 were empty and therefore excluded for further analysis as enough feces had to be left for the proteome analysis and short-chain fatty acid analysis. Of the 52 remaining samples, 23 met the criteria "were still breastfed at the age of 6 months" and "born vaginally." The samples were randomly chosen, and six samples (samples 1-6) were placed in the group "high in *Bacteroides*," and five samples (samples 7-11) were placed in the group "high in *Bifidobacterium*."

3.2. Gut Microbiota Composition in Selected Samples

In the group high in *Bacteroides*, the general abundance of *Bacteroides* was 37.3% and 5.2% of *Bifidobacterium* (figure 3.1). The second most abundant genus in this group was the *Clostridium sensu stricto 1* with 12.66%, followed by *Escherichia-Shigella* with 10.06%. In the group high in *Bifidobacterium*, the general abundance of *Bifidobacterium* was 49.23% and 8.55% of *Bacteroides*. The second most abundant genus in this group was *Escherichia-Shigella*, with 10.43%.

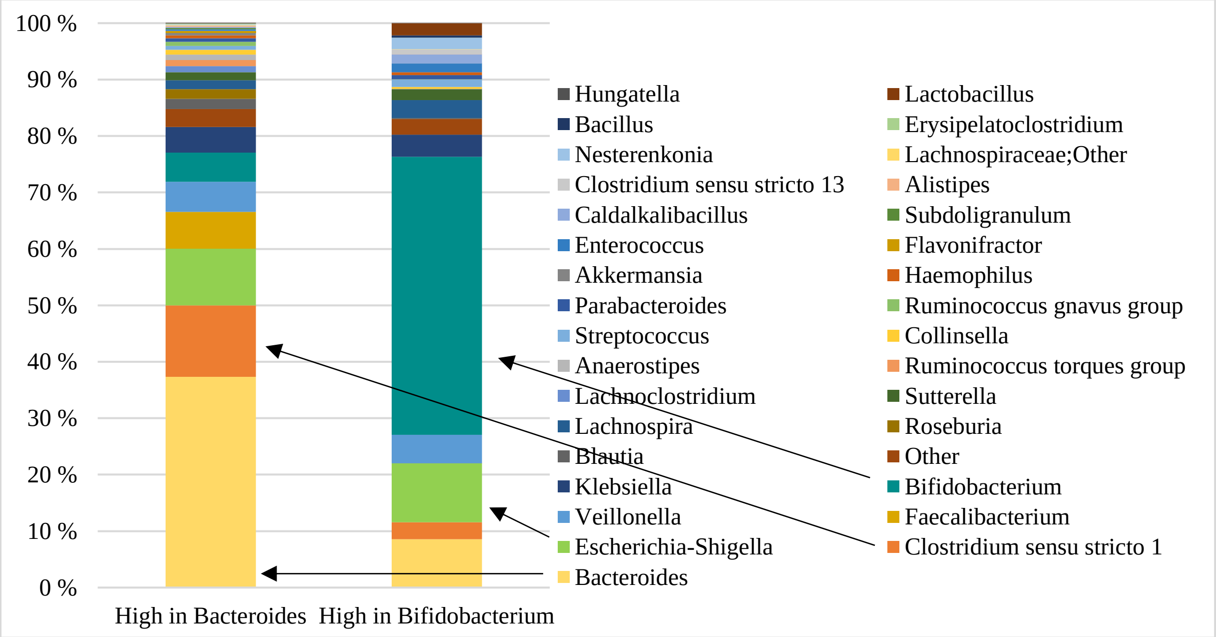

Figure 3.1. Average gut microbiota composition of two groups based on 16S sequencing results. A bar chart representing the average abundance for samples high in *Bacteroides* (1-6) and samples high in *Bifidobacterium* (7-11), given in percentage.

There were great inter-individual variations in the gut microbiota composition between all samples included in the experiments (figure 3.2). The abundance of *Bifidobacterium* and

Bacteroides in the samples was quite varying due to the sample criteria, ranging from not being present at all to being the main genus present. Sample two had the highest abundance of *Bacteroides* with 55.20%, while sample 11 had the highest abundance of *Bifidobacterium* with 68.92%. Sample nine had an approximate ratio of 2:1 of *Bifidobacterium* and *Bacteroides*, respectively, and was the sample with the evenest abundance of the two genera. Additionally, *Clostridium sensu stricto 1* was the main colonizer within sample 3, highly contributing to the high average of this genus in the high in *Bacteroides* group observed in figure 3.1.

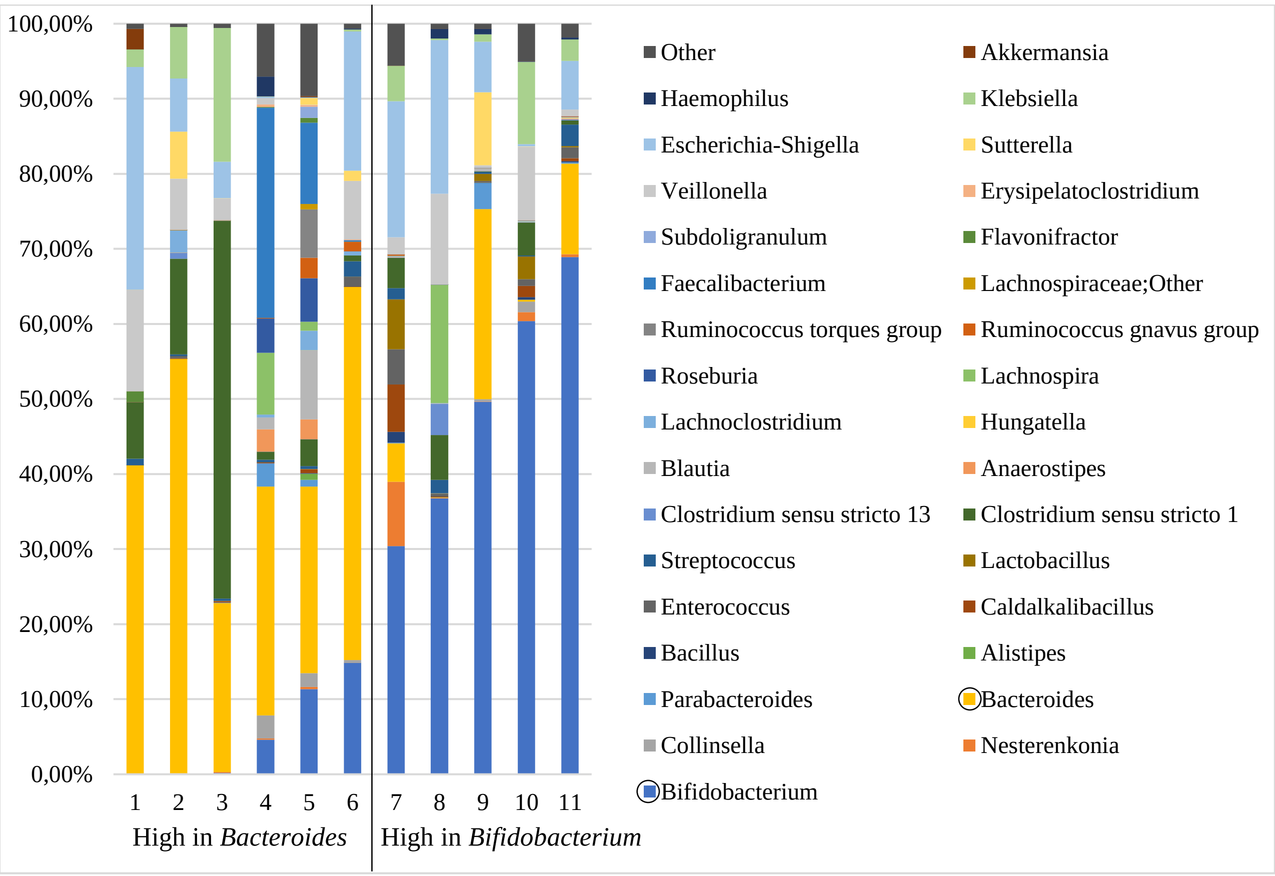

Figure 3.2. Gut microbiota composition samples 1-11 based on 16S sequencing results. The abundance of different genera is given in percentage for samples 1-11, which are samples high in *Bacteroides* (1-6) and samples high in *Bifidobacterium* (7-11). Colour-coded explanations of species to the right, with *Bifidobacterium* and *Bacteroides* marked with circles.

3.3. Composition of *Bacteroides* and *Bifidobacterium* Species from Shotgun Sequencing

Of the 62 complete MAGs, 40 were annotated as *Bacteroides* or *Bifidobacterium*. Two MAGs were annotated as *Parabacteroides*. For samples high in *Bacteroides* (1-6), the total average abundance of the *Bacteroides* genus was 64.44 %, whereas the highest abundance belonged to *Bacteroides fragilis* NCTC 9343 (22.1%) and *Bacteroides dorei* DSM 17855 (15.45%). 2.64% was annotated to the *Bifidobacterium* genus, which included 1.26% of *Bifidobacterium longum* *subsp. infantis* ATC 15697 and 0.51% of *Bifidobacterium longum subsp. longum* JCM 1217. For samples high in *Bifidobacterium* (7-11), the total average abundance of the *Bifidobacterium* genus was 64.65%, with the highest abundance belonging to *Bifidobacterium longum subsp. infantis* (23.77%), *Bifidobacterium bifidum* ATCC 29521 (14.34%), and *Bifidobacterium longum subsp. longum* JCM 1217 (14.08%). A total of 6.77 % was annotated to the *Bacteroides* genus, with most of them belonging to *Bacteroides fragilis* (1.55%) and *Bacteroides fragilis* NCTC 9343 (1.41%). The remaining percentage of the MAGs (uncertain groups) were either unmapped, removed, or not considered pure enough to be annotated to one of the genera. Figure 3.3 represents only the abundance of MAGs from *Bacteroides* (and *Parabacteroides*) and *Bifidobacterium* present in samples 1-11 since all other genera were excluded from the shotgun dataset.

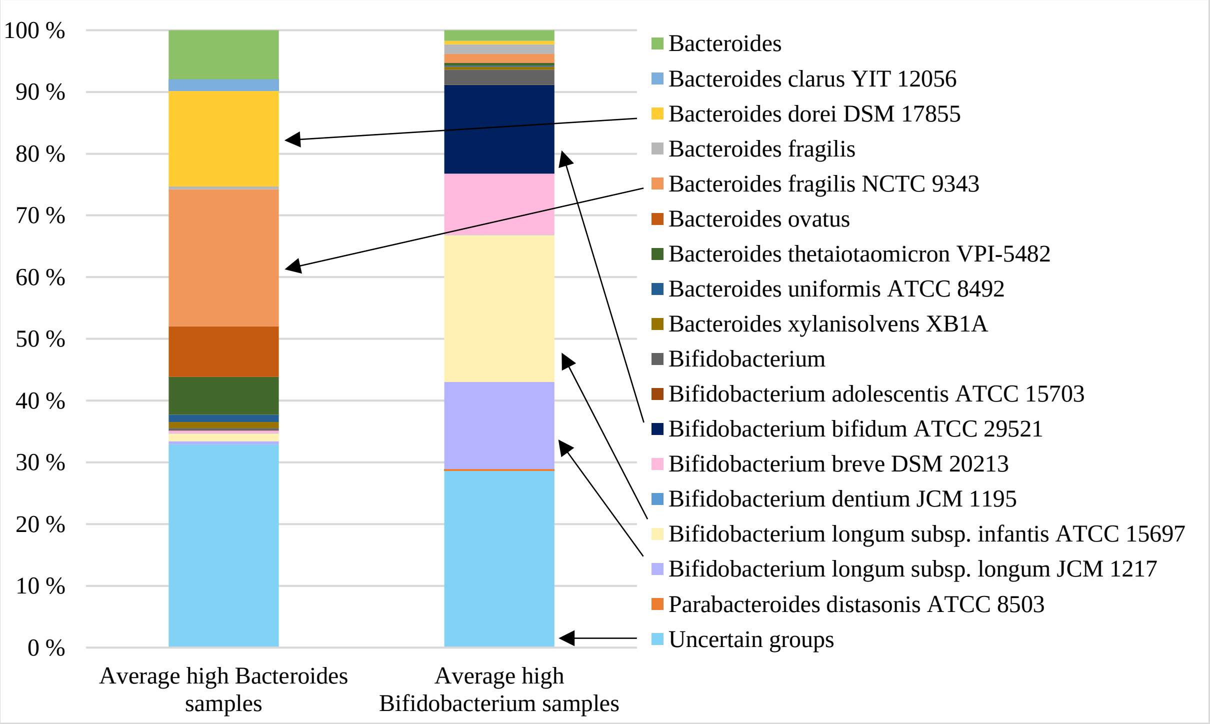

Figure 3.3. Shotgun sequencing results of MAGs annotated to *Bacteroides* (and *Parabacteroides)* and *Bifidobacterium* species. The figure illustrates the average abundance of MAGs annotated as either *Bacteroides*, *Parabacteroides*, or *Bifidobacterium* genus in samples high in *Bacteroides* (1, 2, 3, 6) and samples high in *Bifidobacterium* (7-11) down to a strain level. The category "Bifidobacterium" had several species included in the MAGs and could not be annotated to just one. These included *Bifidobacterium adolescentis, B. angulatum, B. bifidum, B. breve, B.catenulatum, B. dentium, B. longum, B. longum subsp. infantis, B. londum subsp. longum, B. pseudocatenulatum, B. thermophilum, B. ruminantum, B. pullorum subsp. gallinarum, B. pseudolongum subsp. globosum, B. kashiwanohense.* Accordingly, the category "Bacteroides" had several species included in the MAGs and could not be annotated to just one. These included *Phocaeicola,* uncultured *Bacteroides sp., Parabacteroides sp. AN4, Parabacteroides merdae, P. distasonis, P. gordonii, Bacteroides sp. CAG:545, Bacteroides uniformis, B. xylanisolvens, Bacteroides thetaoitamicron, B. stercoris, B. ovatus, B. dorei, B. intestinalis, B. cellulosilyticus, B. bouchesdurhonensis, B. plebeius, B. sartorii, B. coprophilus, B. stercorirosoris, B. finegoldii, B. acidifaciens* and *B. salyersiae.* The category "uncertain groups" are MAGs that were unmapped, removed samples, and MAGs that had different genera annotated to the nodes within the MAGs and were not considered pure.

3.4. Extraction and Identification of Unique Genes for *Bacteroides* and *Bifidobacterium*

After annotation and removing all genes that had e-value > 1e^{-10}, the final number of genes from *Bacteroides* and *Bifidobacterium* was 34 025 out of the total 70 251 genes, removing more than half of the initial database from the shotgun data. The COG categories were then investigated in the database created in Appendix C4. The highest abundance category was “S” (function unknown), with 6560 out of 34 025 genes. The second most abundant COG category was “G” (Carbohydrate transport and metabolism), with 4023 genes.

3.5. Proteome Analysis

Filtering protein data

After mass spectrometry and the raw data treatment were performed, a total of 1541 unique proteins were mapped to the database. After the filtration performed in Perseus, including removal of proteins only identified by site, reverse proteins, potential contaminants, human proteins, and all proteins that only appeared once in the dataset, the result was a table containing 717 unique proteins. These proteins were used further in the analysis.

3.5.1. Calculations and Clustering of Proteins with Perseus

All samples were run in parallels throughout the protein extraction protocol, meaning each fecal sample had a replicate. In Perseus, the Pearson correlation between replicates was calculated and ranged between 0.619-0.940 (table 3.1). Sample 3 had no correlation due to no results in replicate 3b after MS.

Table 3.1. Pearson correlation between replicates. The Pearson correlation was calculated by Perseus. The table illustrates the correlation between replicate a and b in samples 1-11. Sample 3b had no results after MS and the pearson correlation between 3a and 3b was therefore 0.

Sample	1a/1b	2a/2b	3a/3b	4a/4b	5a/5b	6a/6b	7a/7a	8a/8b	9a/9b	10a/10b	11a/11b
Pearson correlation	0.925	0.619	0	0.827	0.865	0.693	0.940	0.83	0.633	0.835	0.801

The correlations indicated a similar outcome of unique proteins from each sample-replicate. The replicates would also cluster together when clustering based on LFQ intensity and unique proteins. Therefore, all replicates were merged. The average LFQ intensity between replicates were calculated, including the total LFQ intensity in each sample (table 3.2). Samples 3, 4, 8, and 9 were removed due to low total LFQ-intensity (< 1000), which can also be observed in the histograms in Appendix A2, Figure 3. Sample 1 and 5 stood out with a total LFQ intensity > 5000. The other samples had a total LFQ intensity laying around 2000.

Table 3.2. Total LFQ intensity in each sample. The LFQ intensity was measured by mass spectrometry. The represented LFQ intensity is after the protein-filtration performed in Perseus and calculating the average between replicates.

Sample	1	2	3	4	5	6	7	8	9	10	11
Total LFQ-intensity	5434.78	1405.38	173.86	369.62	5579.33	2200.23	2224.58	628.65	757.84	1841.25	2145.83

Sample 1 and 5 had big differences between them but also differed from the other samples. There was a high presence of proteins in samples 1 and 5 that were hardly present in the other samples, especially the proteins in clusters 3 and 4 (figure 3.4A).

The annotation in Perseus revealed which genus was prevalent in each cluster (figure 3.4B). Cluster 1 had proteins only belonging to the *Bacteroides* species, while cluster 2 had proteins only annotated to the *Bifidobacterium* species, except for one protein that was annotated to the *Bacteroides* genus. Cluster 3 had proteins from both genera but primarily from *Bacteroides,* with 216 out of 254 proteins. Cluster 4 had proteins almost exclusive from the *Bacteroides* species, while clusters 5 and 6 had proteins that belonged only to the *Bifidobacterium* species.

Almost 85% of the proteins in sample 1 belonged to cluster 4, and around 11% of the proteins belonged to cluster 1 (figure 3.4C). Both cluster 1 and 4 were *Bacteroides* dominant clusters. Cluster 3, which was primarily proteins from species of *Bacteroides*, was prevalent within the high in *Bacteroides* group and dominated samples 2, 5, and 6. It was also quite abundant in samples 10 and 11 in the high in *Bifidobacterium* group, although these samples had a more even distribution of all 6 clusters. Sample 7 had over 70% of its proteins from cluster 2, while sample 11 had approximately 20% of its proteins from cluster 2, which was a *Bifidobacterium* dominated cluster. Both samples belonged to the high in *Bifidobacterium* group. Cluster 6 appears mainly in samples high in *Bifidobacterium* but is observed to some extent in samples high in *Bacteroides*, like sample 5 and 6.

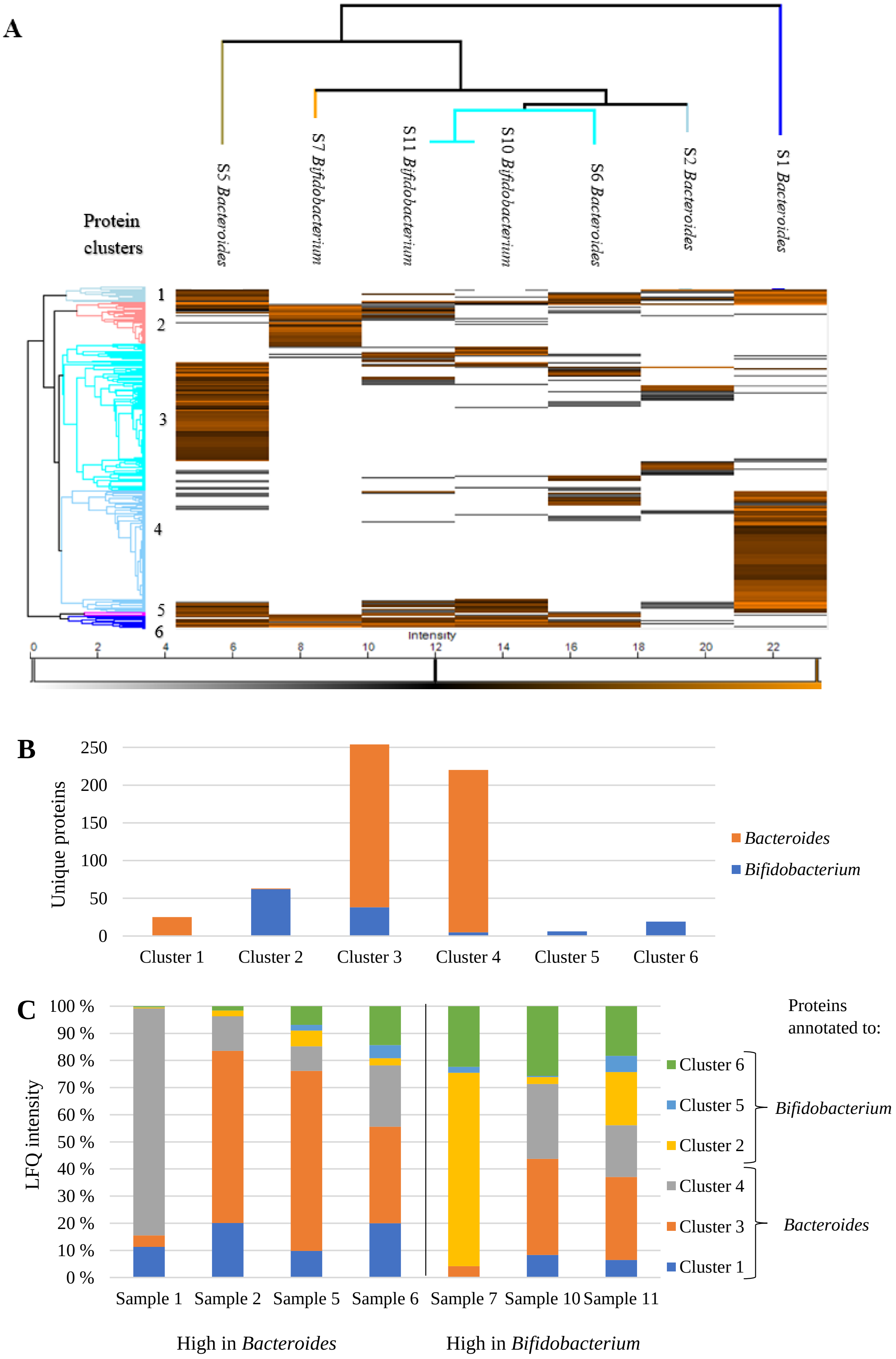
A
S5 Bacteroides
S7 Bifidobacterium
S11 Bifidobacterium
S10 Bifidobacterium
S6 Bacteroides
S2 Bacteroides
S1 Bacteroides
Protein clusters
1
2
3
4
5
6
Intensity
0
2
4
6
8
10
12
14
16
18
20
22
B
Unique proteins
250
200
150
100
50
0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Bacteroides
Bifidobacterium
C
LFQ intensity
100 %
90 %
80 %
70 %
60 %
50 %
40 %
30 %
20 %
10 %
0 %
Sample 1
Sample 2
Sample 5
Sample 6
Sample 7
Sample 10
Sample 11
High in Bacteroides
High in Bifidobacterium
Proteins annotated to:
Cluster 6
Cluster 5
Cluster 2
Bifidobacterium
Cluster 4
Cluster 3
Cluster 1
Bacteroides

Figure 3.4. Clustering of proteins and abundance in samples high in *Bacteroides* and high in *Bifidobacterium*.
(A)Heatmap of the average LFQ intensity in analyzed samples. All samples with higher total LFQ intensity than 1000 and the proteins that appeared in clusters 1-6. S1, S2, S5, and S6 are samples from infants who had high abundance of Bacteroides, while S7, S10, and S11 are samples from infants who had high abundance *Bifidobacterium.* The figure was made in Perseus version 1.6.6.0.
(B)Bar chart of unique proteins in all six clusters. Provides the count of unique proteins found in all given clusters and to which genus they were annotated. The bar chart illustrates which genus is most abundant in each protein cluster.
(C)Bar chart of clusters that are present in all samples. The quantity of proteins given in LFQ intensity illustrates the abundance of the protein clusters present in the samples high in *Bacteroides* and samples high in *Bifidobacterium*.

3.5.2. Glycoside Hydrolases

Shotgun data

After the annotation performed in dbCAN, all glycoside hydrolases were annotated to the proteins from the MS and shotgun data separately. In the shotgun data, all GH-families related to HMO- and mucin-degradation were found in at least one node that came from *Bifidobacterium* (Table 3.3). *Bacteroides* had 10 out of 15 GH-families that were found in at least two nodes. The genus seems to miss the HMO-related GH112 and GH1, the mucin-related GH101 and GH129, and the HMO- and mucin related GH85.

The most abundant GHs detected in both genera were the GHs related to both HMO- and Mucin degradation, like the GH families GH2 and GH42 (β-galactosidases), GH20 family (β-hexosaminidases), and GH29 family (fucosidases). The GH33 (sialidases) were also found in both genera, yet not as abundant as the other. Additionally, *Bifidobacterium* had several potential proteins from both GH112 and GH1 in the shotgun data, whereas *Bacteroides* did not have any of these GH families.

Protein data

Out of the proteins observed in the proteomics, five of 15 GH-families were identified from *Bifidobacterium,* and three were identified from *Bacteroides* (table 3.3). The fucosidase in GH29 was a common GH for the two genera. GHs found in *Bacteroides* was the HMO-related GH18 which is an endo- β-N-acetylglucosaminidase, and the HMO- and Mucin-related GH20 (Lacto-N-biosidase and β-hexosaminidase), including the GH29 (fucosidase). GHs found in *Bifidobacterium* were the HMO-related GH112 (GLNBP) and GH136 (Lacto-N-biosidase). Additionally , the HMO-and mucin-related GH95 and GH29 (fucosidases), and GH2 (β-galactosidase). The most abundant GH in the protein data was the β-galactosidase from the GH2 family, which was only found in *Bifidobacterium*.

Table 3.3. **Glycoside hydrolases present in the shotgun data and proteomics data.** Heatmap illustrating the presence of potential proteins and proteins detected in mass spectrometry within the genomes of the genera *Bacteroides* and *Bifidobacterium*. Complementary information about the GH families can be found in table 1.1 in the introduction. The table was made based on the information in Ioannou et al. (Ioannou et al., 2021) and Tailford et al. (Tailford et al., 2015).

	GH Family	Present in shotgun data		Present as proteins	
		Bacteroides	*Bifidobacterium*	*Bacteroides*	*Bifidobacterium*
HMO related	GH18	8	1	1	0
	GH112	0	10	0	1
	GH136	2	3	0	1
	GH1	0	8	0	0
	GH35	8	3	0	0
Mucin related	GH89	5	1	0	0
	GH101	0	6	0	0
	GH129	0	5	0	0
HMO and mucin related	GH95	7	6	0	2
	GH2	65	37	0	5
	GH42	3	27	0	0
	GH20	29	12	1	0
	GH29	16	8	2	2
	GH33	2	8	0	0
	GH85	0	2	0	0

3.6. KEGG Mapper Reconstruct – Metabolic Pathways

By using the K-numbers withdrawn from the EggNOG annotation database, KEGG Mapper Reconstruct was used to visualize the pathways that were thought to be common in *Bacteroides* and *Bifidobacterium*. From the degradation of HMOs, galactose is a common monosaccharide that most organisms can further utilize (Raimondi et al., 2021). Therefore, a reconstruction of the galactose metabolism and Leloir pathway was done (figure 3.5A). *Bifidobacterium* may also use the GNB/LNB-pathway (figure 3.5B). As most genera in the gut microbiota produce acetate, this pathway was reconstructed (figure 3.5C), and because *Bacteroides* is recognized for utilizing propionate through the succinate pathway, this pathway was reconstructed as well (figure 3.5D).

The pathway of the galactose metabolism was complete for the *Bifidobacterium* genus, apart from GalM (EC 5.1.3.3), which only appeared in the genome and proteome of *Bacteroides* (figure 3.5A). The other proteins used in the Leloir pathway were present in *Bifidobacterium*, with GalE (EC 5.1.3.2) and GalT1 (EC 2.7.7.12) detected in the proteomics data and shotgun data. *Bacteroides* did not have the complete Leloir pathway, as the GalT1 was absent. The transferase (EC 2.7.7.9) in the galactose metabolism that converts the UDP-glucose into glucose-1P was also absent from the datasets for *Bacteroides*.

The LNB/GNB pathway was complete within the *Bifidobacterium* genus but not in *Bacteroides* (figure 3.5B). *Bacteroides* lacked the GLNBP enzyme belonging to family GH112, that removes the LacNAc and GalNAc from the reducing ends of HMOs and mucins. The absence is also observed in Table 3.3. Additionally, *Bacteroides* lacked the NahK (EC 2.7.1.162) and GalT2 (EC 2.7.7.12). All proteins in this pathway were found in the proteomics and shotgun database for *Bifidobacterium* except for GalT2 (EC 2.7.7.12) which was only observed in the shotgun data.

The transformation of phosphoenolpyruvate and pyruvate into acetate was complete for both genera, only illustrated by different pathways (figure 3.5C). *Bifidobacterium* illustrated the potential of producing lactate (EC 1.1.1.27), formate (EC 2.3.1.54), and ethanol (EC 1.1.1.1/2) on the way to acetate. *Bacteroides,* on the other hand, had the enzyme pyruvate dehydrogenase (EC 1.2.5.1) to make acetate out of pyruvate directly in both the shotgun and proteomic data, which *Bifidobacterium* was lacking. The potential of producing acetate through Acetyl-CoA was shown in both genera, though *Bacteroides* had more enzymes present, including the Acetyl-CoA-hydrolase (EC 3.1.2.1), which was also discovered in the proteomics.

Three pathways exist to produce propionate: the acrylate pathway, the propanediol pathway, and the succinate pathway. *Bacteroides* illustrated the potential for the complete succinate pathway, and *Bifidobacterium* did not (figure 3.5D). No proteins in this pathway were detected in the proteomics data, only in the shotgun database. The acrylate pathway and the propanediol pathway were also investigated, but neither of the two genera had complete pathways for propionate production through these pathways.

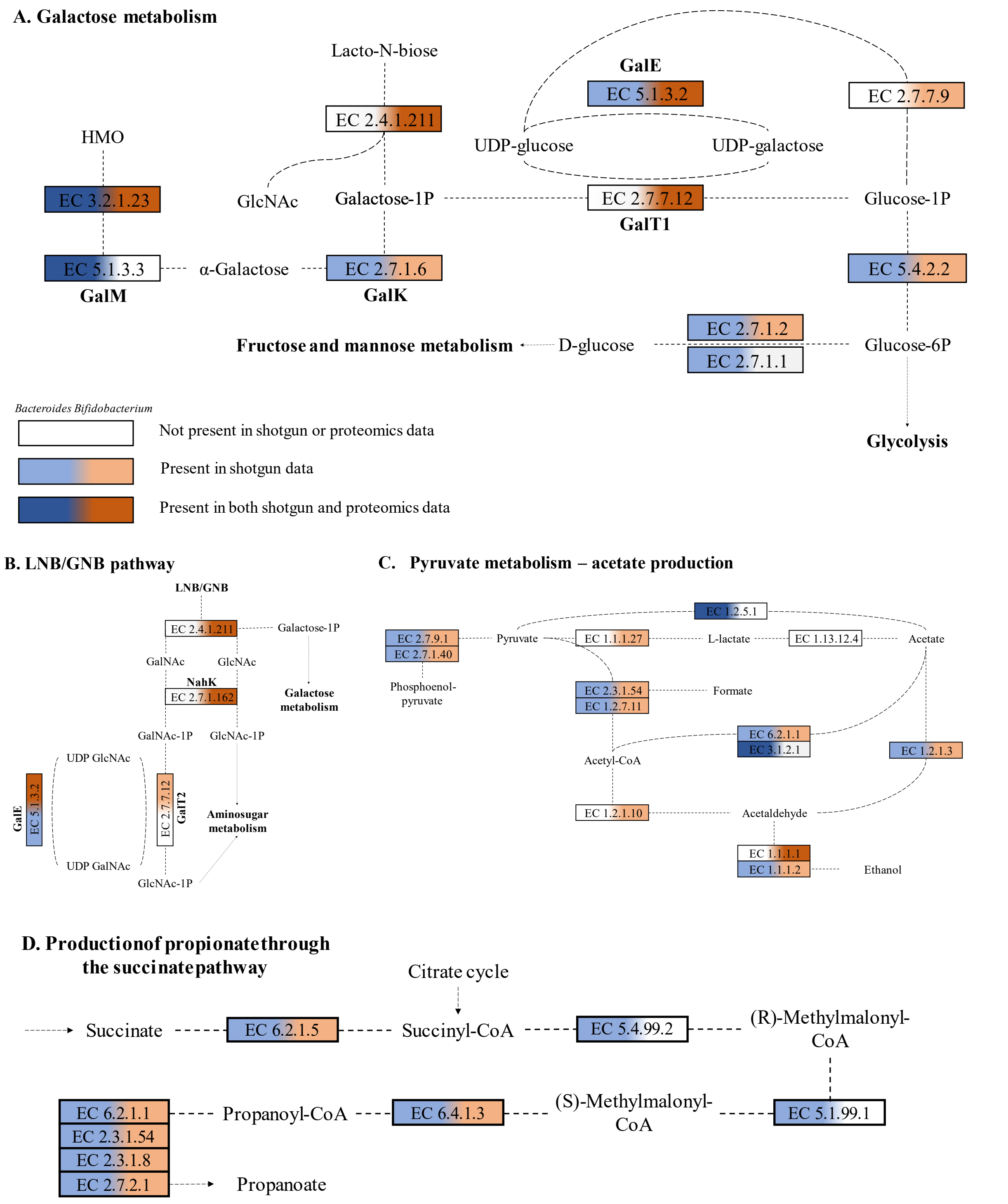

Figure 3.5. Metabolic pathways. Pathways were investigated using the potential proteins in shotgun data and the proteins detected in the proteome analysis. White indicates that the protein was not present in either dataset. The bright colors indicate the presence in shotgun data, while darker colors indicate the presence in both shotgun data and proteomics data. Blue colors belong to *Bacteroides*, and orange colors belong to *Bifidobacterium*. This is also illustrated in (A).
(A) Galactose metabolism. The galactose utilization pathway from HMOs and GNB/LNB to galactose is further processed into glucose and used in glycolysis and other energy-conserving metabolisms. The Leloir pathway is integrated into the galactose metabolism from the utilization of galactose into glucose (GalMKTE).
(B) GNB/LNB pathway. With enzymes and substrates involved.
(C) Pyruvate metabolism. With the production of acetate and the enzymes and substrates involved.
(D) Succinate pathway. Illustrates the production of propionate and the enzymes and substrates involved.

Finally, because of the mucin-utilizing species to an extent being dependent on sulfatases, enzymes with EC number 3.1.6.- were searched for in the shotgun database. The arylsulfatase (EC 3.1.6.1) and choline-sulfatase (EC 3.1.6.6) were observed in the *Bacteroides* genus and not in the *Bifidobacterium* genus. The proteins were not detected in the proteome analysis.

3.7. Short-Chain Fatty Acid analysis

There were great inter-individual variations between the samples when investigating SCFA levels. Samples 2-4 had higher butyrate levels than the others (figure 3.6). Acetate was the SCFA with the highest levels in all samples, particularly in sample 11, with a concentration of 96.1 mmol/kg feces of the total SCFAs, which were 126.44 mmol/kg feces. The ratio of acetate to propionate in sample 1 and 7 were almost 1:1.

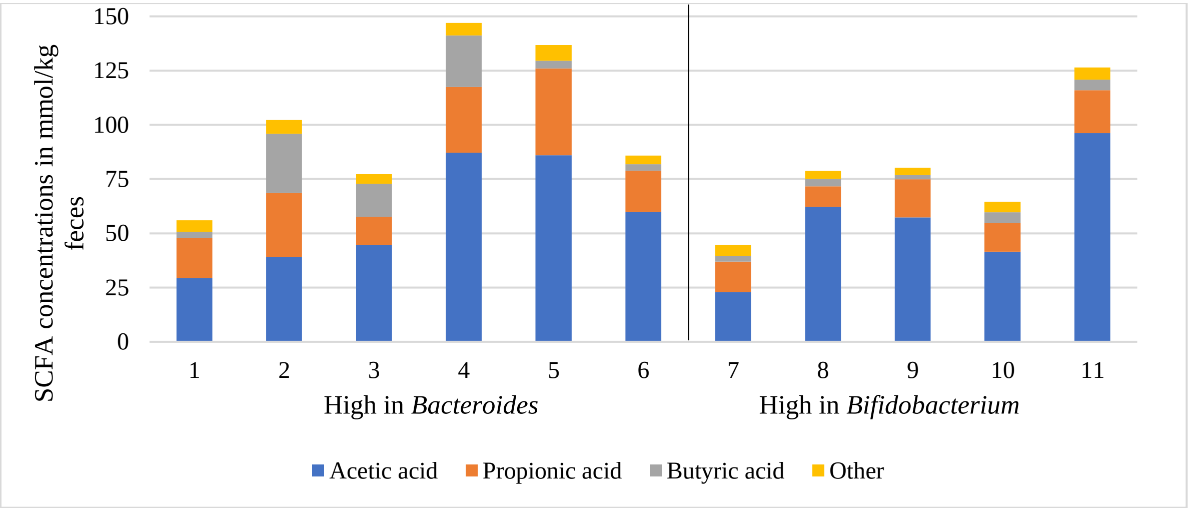

Figure 3.6. SCFA levels in samples 1-11, given in mmol/kg feces. A bar chart illustrating the absolute values of the different SCFA-concentrations in samples high in *Bacteroides* (1-6) and samples high in *Bifidobacterium* (7-11).

The overall levels of SCFAs detected were higher in the high in *Bacteroides* group compared to the *Bifidobacterium* group, with a total average of SCFA-concentration at 100.82 mmol/kg feces (figure 3.7). For comparison, the high in *Bifidobacterium* group had a total average of 78.91 mmol/kg feces with SCFAs. Acetate levels were approximately equal, but when applying relative values, acetate accounted for 70% of the total SCFA levels in the high in *Bifidobacterium* group and 57% of the total SCFA levels in the *Bacteroides* group. Propionate- and butyrate-levels were substantially higher in the *Bacteroides* group. Especially butyrate, with almost three times higher levels than what was observed in the *Bifidobacterium* group.

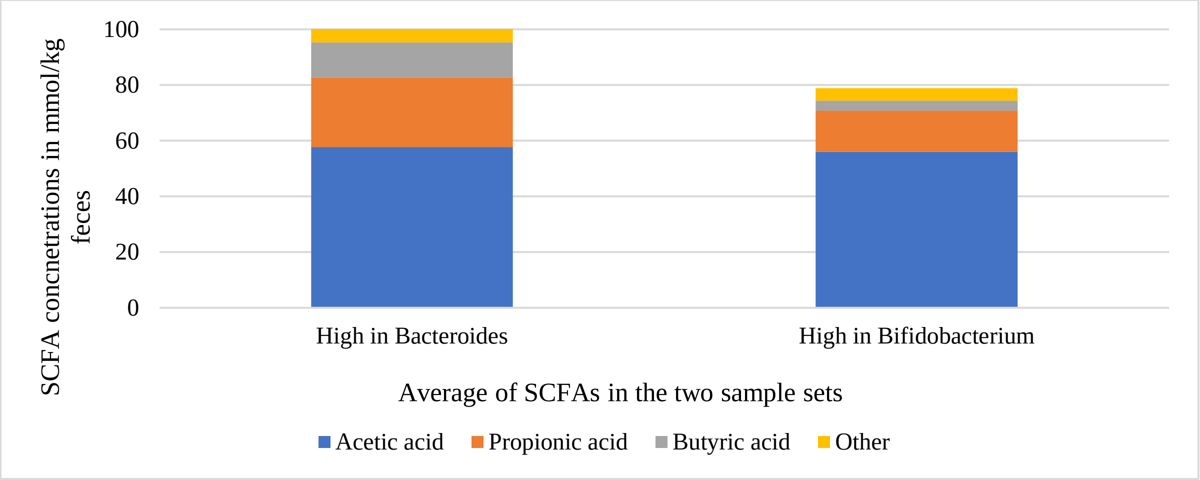

Figure 3.7. Bar chart representing the average of SCFA levels in two sample sets. The bar chart illustrates the difference in SCFA production between the samples high in *Bacteroides* (1-6) and the samples high in *Bifidobacterium* (7-11) with absolute values.

When the correlations performed in the statistical analysis were investigated, no positive correlations between *Bacteroides* or *Bifidobacterium* and the SCFAs were observed (Appendix C7, figure C7.1). Only a weak negative correlation was observed for *Bifidobacterium* and butyrate with a p-value < 0.1. Lastly, when splitting up the two groups in separate correlation plots (p-value < 0.05), still no correlations were observed between the two genera and the SCFAs.

4. Discussion

Pathways and Glycoside Hydrolases present in the genomics and proteomics of *Bifidobacterium* and *Bacteroides* genus.'

4.1. The Leloir- and LNB/GNB- Pathways – Enzymes in the Galactose Metabolism

The *Bifidobacterium* genus had a complete presentation of all GHs related to HMO- and Mucin-utilization in the intestine in the gene set provided from the shotgun sequencing data (Table 3.3). All genes were also present in the galactose metabolism (figure 3.5A), except for the galactose mutarotase, GalM (EC 5.1.3.3), which converts β-Gal into α-Gal. α-Gal is the substrate for galactokinase (EC 2.7.1.6), and the GalM is needed to convert the galactose into the correct form to enter the Leloir pathway. However, all other enzymes were present in the shotgun data, including GalE and GalT1 in the proteomics. The genus may just convert α-Gal and LNB into Galactose-1P for the continuing metabolism of the Leloir pathway without being dependent on GalM. The conversion of β-Gal into α-Gal could also be performed by other genera, like the *Bacteroides* genus, which had the GalM detected both in the shotgun and proteomic data. Another study investigating the gene set of *B. longum* NCC2705 found no gene cluster for the full set of enzymes in the Leloir pathway, only a cluster containing the GalK and GalT (Nishimoto & Kitaoka, 2007). Little reports were found about the GalM and the *Bifidobacterium* genus. However, the genes encoding the Leloir enzymes seem to be observed in other species (Nishimoto & Kitaoka, 2007).

Nevertheless, *Bifidobacterium* has an alternative way of degrading galactose through the LNB/GNB pathway (figure 3.5B). All proteins were observed in the shotgun data of this pathway, and only GalT2 was lacking from the proteomics. It is likely that this is a preferred pathway for *Bifidobacterium* and that it does not depend on the GalM. Due to the direct use of Gal1P, it will not require the GalK (EC 2.7.1.6) either, making it less energy-requiring and thus more appealing for the *Bifidobacterium* genus (De Bruyn et al., 2013). Finally, the sequence alignments of GalT1 and GalT2 revealed a presence of both, whereas the coexistence probably is due to the LNB/GNB pathway, which has been suggested to be present only in *B. bifidum, B. longum* subsp. *infantis, B longum* subsp. *longum,* and *B. breve* strains of *Bifidobacterium* (De Bruyn et al., 2013; Turroni et al., 2010). As all these genomes were represented as MAGs (figure 3.3) used in the shotgun data, the presence of the GalT proteins was expected. However, GalT was not observed in either the shotgun data or the proteomics data for *Bacteroides.* This could suggest that *Bacteroides* does not have the full capacity of degrading galactose but partially and further rely on other species like the *Bifidobacterium* species, for instance, to convert Gal1P to Glc1P.

For comparison, the *Bacteroides* genus lacked 5 out of the 15 GHs in the gene set provided from the shotgun sequencing data where *Bifidobacterium* had all (table 3.3). Among these were the HMO-related GH112. The GH112 family holds the lacto-N-biose phosphorylase or galacto-N-biose phosphorylase (GLNBP) (EC 2.4.1.211) enzyme. It utilizes galactosyl-β1,3-N-acetylhexosamines (LNB or GNB) and degrades them into Galactose-1P and N-acetylhexosamines which is GalNAc or GlcNAc (Ioannou et al., 2021). Since this pathway is primarily known to be found in species of the *Bifidobacterium* genus, it was not expected to be present in the genome of *Bacteroides* (figure 3.5B). However, *Bacteroides* utilize GNB in mucins and must use other mechanisms and enzymes than *Bifidobacterium*. The GH112 is specific towards the β1,3-linkages of LNB and GNB in HMOs and mucins, but the N-acetylglucosaminidases of families GH18 and GH85 target β1,3/4-linkages of LNB and LacNAc, whereas the β1,4-linkages are highly abundant in mucins (Bell & Juge, 2021). GH18, which is related to HMO degradation, was found in the gene set of *Bacteroides* with eight GHs, while *Bifidobacterium* had one (Table 3.3). The GH18 was also found in the proteomics data of *Bacteroides,* as one of three GHs discovered here. GH85, which is both HMO and mucin related, was not present within the *Bacteroides* shotgun data, but two GH85 were found within the *Bifidobacterium* genus. Either way, the presence of the GH18 may suggest that even if the *Bacteroides* genus lacks certain HMO-related GHs such as the GH112, it does not mean it lacks the capability of degrading β1,3/4-linkages of LNB.

4.2. β-Hexosaminidase Activity needed for HMO- and Mucin Degradation

The β-hexosaminidase belonging to the GH20 family (EC 3.2.1.52) targets the β1-3 or β1-6 linkages between GlcNAc and Gal, releasing the adjacent LNB. These linkages are found in all HMO structures and mucin, and therefore, the enzyme is necessary to degrade both substrates. It appears 29 times in the *Bacteroides* genus and 12 times in the *Bifidobacterium* genus in the shotgun data (Table 3.3). One GH20 was found annotated to *Bacteroides* in the proteomic data but not *Bifidobacterium*. This means that the *Bacteroides* express the gene and produce the protein needed to cleave the β1-3/6 linkages in HMO and mucin, while *Bifidobacterium* has the potential to produce the protein.

In addition to the β-hexosaminidase, a Lacto-N-biosidase (EC 3.2.1.140) in the GH20 family targets the same linkages, releasing lactose from the reducing ends of HMOs (Ioannou et al., 2021). There is an additional lacto-N-biosidase in family GH136 (EC 3.2.1.140), but it only targets the β1-3 linkage in LNB, resulting in a smaller range for this enzyme than the GH20 Lacto-N-biosidase. The GH136 was also found in the genomes of both genera, including one protein in the proteomics of *Bifidobacterium*. Finally, this illustrates the ability of both genera to release lactose as the presence of GH20 and GH136 was present in the shotgun and proteomics data. The expression of GH20 in *B. thetaoitamicron* was upregulated when grown on HMO carbon sources in Marcobal et al. (Marcobal et al., 2011), supporting the presence of the GH in *Bacteroides* in the database and its potential ability to degrade the respective linkages in HMOs.

4.3. Degrading Lactose

The release of lactose from the reducing end of HMOs performed by β-hexosaminidases enables it for β-galactosidases (EC 3.2.1.23) that can degrade β-1,4-linkages between Gal and Glc in lactose. β-galactosidases that degrade these linkages can be found in the families of GH1, GH2, GH35, and GH42. According to Ioannou et al. (Ioannou et al., 2021), this enzyme has not been characterized in the GH1 family from the highly abundant bacterial species in the infant gut. However, glycosidase profiles were characterized by GH1 β-galactosidases in MAGs of *B. breve* in a machine learning approach study performed by Sabater et al. (Sabater et al., 2021). The GH1 was observed eight times in the shotgun data for the *Bifidobacterium* genus (table 3.3), which may indicate a potential for the genus to use the galactosidase in HMO-degradation. However, it was not discovered in the proteomics of the genus. The GH1 was not present in the data for *Bacteroides*, but the GH2, GH35, and GH42 were highly present, whereas the GH2 appeared 65 times. It is related to both HMO- and mucin degradation, and the high occurrence could support the suggestion of *Bacteroides* activating the same genes for HMO degradation as for the utilization of host mucus glycans, which was proposed by Marcobal et al. (Marcobal et al., 2011). No β-galactosidases were detected in the proteomics for *Bacteroides*, but the genus had the gene and potential ability to express it and produce the protein.

The β-galactosidases do not only degrade lactose (Miwa et al., 2010). They may cleave general β-linkages that yield galactose but with varying specificity. The GH42 seems to have broader specificity than the other galactosidases and may degrade β1,3/6-linkages as well as the β1,4-linkages. This explains the relation to mucins, which do not appear to have Glc, but GlcNAc and GalNAc, to mention some. As with *Bacteroides*, the GH2 was also highly abundant in *Bifidobacterium*, with 37 proteins detected in the shotgun data, including five proteins detected in the proteomics. The second most abundant galactosidase was the GH42, with 27 proteins. Finally, the findings suggest that *Bifidobacterium* not only specializes in HMO degradation but also utilize mucin. Accordingly, it supports the possibility that *Bacteroides* might do the same.

4.4. Fucosidases – the Removal of Fucose from HMOs and Mucin

Fucose is attached with α1-2/3/4 linkages on HMOs and Mucins, and these are removed with fucosidases which were highly abundant in both the *Bacteroides* and *Bifidobacterium* genera. The fucosidases of the GH95 family cleave the α1-2 linkage. Seven of this GH were observed in the shotgun data of *Bacteroides,* while six were observed for *Bifidobacterium*. The latter also had two GHs present as proteins (Table 3.3). The GH29 cleaves the α1-3/4 linkages, and eight and 16 GHs were found in the shotgun data for *Bifidobacterium* and *Bacteroides,* respectively. The GH was also found in the proteomic data for both genera. The presence of these in each genus illustrates their potential for utilizing mucins and HMOs with fucose attached.

4.5. Sialidases – Sialic Acid as a Potential Energy Source

Sialidases are present in the GH33 family, and they are important for removing sialic acids in HMOs and mucins, as they cleave the α2-3/6 linkage between Neu5Ac and the saccharide. No proteins of this GH family were found for either genus, but they appeared in the shotgun data for both, with two and eight GHs in *Bacteroides* and *Bifidobacterium*, respectively (Table 3.3). From the shotgun data, there is potential for both *Bacteroides* and *Bifidobacterium* genus to remove sialic acids using sialidases of GH33. Further degradation of the sialic acids can be performed using genes in so-called *nan* gene clusters (Bell et al., 2019). In *Bacteroides fragilis,* a pathway for sialic acid metabolism has been described, converting the Neu5Ac into GlcNAc-6-P (Bell et al., 2019). The latter can be converted to fructose-6-P, which is used in the glycolysis. However, the complete operon was not discovered in the shotgun database, but the pathway could be noteworthy for investigation in later research. The sialic acids must be cleaved off by sialidases to make them available for sialic acid degrading proteins, and the bacteria that harbor these genes are normal habitants of the human gut. As sialic acids are abundant on mucin and certain HMOs, this could be a potential energy source and promote growth for the respective species.

4.6. Mucin Related Glycoside Hydrolases

The O-glycosidic linkages in the protein core of mucins must also be degraded to utilize the mucin oligosaccharide fully. This can be done by enzymes known as endo-α-GalNAcases of the GH101 or GH129 family, which frees the glycan from the protein by cleaving the linkage between galactosylβ1,3-N-acetyl-galactosamine (Gal-β1,3-GalNAc) and Serine or threonine. This linkage is, for instance, found in the abundant core-1 of O-glycan mucins (Koutsioulis et al., 2008). The GalNAcase has been found in *Bifidobacterium longum* and is abundant among *Bifidobacterium* species (Bell & Juge, 2021; Fujita et al., 2005). As *Bacteroides* is adapted to mucin degradation, it would be interesting to observe the presence of the families of GH101 and GH129, but no such presence was detected (Table 3.3). No reports of GH101 or GH129 present in *Bacteroides* species were found either. As expected, the shotgun data provided six of the GH101 family and 5 of the GH129 family in Bifidobacterium. Even if *Bacteroides* is adapted to mucin degradation, it seems to lack endo-α-GalNAcases within the GH101 and GH129 families and will most likely use other mechanisms or glycosidases that were not investigated in this study.

4.7. Sulfatases - Necessary to an Extent for Growth of Gut Microbes

Sulfate is known to be attached to mucin oligosaccharides, capping the glycan, and making it unavailable for further degradation by the bacteria that utilize mucin (Luis et al., 2021). Literature that reports sulfated HMOs, however, is lacking. Kostopoulos et al. stated in 2020 that there are no reports of HMO being sulfated like mucin glycans, but in the same year, Quin et al. stated the identification of 16 sulfonated HMOs for the first time (Kostopoulos et al., 2020; Quin et al., 2020). Mucin utilizing species are dependent on sulfatases to some extent for growth, and the presence of sulfate on HMOs should be further investigated as this topic has little knowledge. One sulfatase in *Bacteroides thetaiotamicron* degrades sulfate on the 3S-Gal of mucins and seems to be of great importance for this species and gut colonization (Luis et al., 2021). Investigating sulfatases in the shotgun data, the arylsulfatase (EC 3.1.6.1) appeared in the genomes of *Bacteroides*. Arylsulfatases catalyze the hydrolysis of aromatic sulfate esters, and Kostopoulos et al. discovered a high expression of these in human milk, as they are part of glycosaminoglycans (GAGs) (Kostopoulos et al., 2020). GAGs are polysaccharides consisting of disaccharides with GlcNAc and GalNAc, and they are highly sulfated, potentially making them an alternative substrate for *Bacteroides* in human milk other than HMOs (Kostopoulos et al., 2020). Therefore, the potential of other components in human milk to be alternatively used as substrates for gut bacteria should be further investigated.

4.8. Short-Chain Fatty Acids

4.8.1. Acetate Production through Pyruvate Metabolism

Both *Bacteroides* and *Bifidobacterium* illustrated complete pathways for acetate production (figure 3.5C). This was expected as most species produce acetate in the intestine, including the species of *Bacteroides* and *Bifidobacterium* (Louis et al., 2007). Investigating the shotgun data, they seem to use different pathways. Both genera reveal the potential of producing acetate through Acetyl-CoA, but only *Bifidobacterium* seems to produce acetate through Acetyl-CoA and acetaldehyde. This makes it possible for the genus to produce ethanol through the pyruvate metabolism. *Bacteroides* had the presence of pyruvate dehydrogenase (EC 1.2.5.1) that directly converts pyruvate into acetate, which was also found in the proteomic data. The protein was not discovered in the data for *Bifidobacterium*. The gene used to produce lactate from pyruvate (EC 1.1.1.27) was only found in the shotgun data of *Bifidobacterium* and concurred with findings in Louis et al. (Louis et al., 2007). Even if most proteins lacked in the proteomic data, both genera illustrated the potential of producing acetate through more than one pathway.

When exploring SCFA levels of the two groups (figure 3.7), the high in *Bifidobacterium* group had higher relative acetate levels than the high in *Bacteroides* group. This could result from several factors but may indicate an imbalance related to more acetate-producing bacteria and less propionate- and butyrate-producing bacteria in the high in *Bifidobacterium* group. Endwise, correlations between *Bacteroides* or *Bifidobacterium* and acetate were not discovered in any correlation analysis performed. This could be explained by the wide range of bacteria that produce this metabolite.

4.8.2. Propionate Production through the Succinate Pathway

Bacteroides revealed the complete pathway for propionate production through the succinate pathway (figure 3.5D). *Bifidobacterium* lacked enzymes converting the Succinyl-CoA into (R)-methylmalonyl-CoA (EC 5.4.99.2) and further into (S)-methylmalonyl-CoA (EC 5.1.99.1). Proteins belonging to the succinate pathway were not detected, although it does not reject the pathway from the metabolism of *Bacteroides,* which are known to produce propionate (Rios-Covian et al., 2017). There is a possibility that the proteins were not yet produced when extracting proteins from the samples. The ability of *Bifidobacterium* to produce propionate is uncertain, although when searching for literature, it seems like certain *Bifidobacterium* species produce 1,2-propanediol. 1,2-propanediol is a precursor for propionate production, and it seems the genus promotes the growth of species that produce propionate through that pathway (Bunesova et al., 2018). Lactate, which was found to be produced by *Bifidobacterium* and not *Bacteroides,* is a precursor in propionate production through the acrylate pathway. Therefore, the alternative pathways for propionate production were investigated – but neither of the genera had complete sets of enzymes through the acrylate pathway or the propanediol pathway. Accordingly, this study did not provide any indications of *Bifidobacterium* producing propionate. Finally, the propionate levels were higher for the group high in *Bacteroides* (figure 3.7). This could be explained by the sample criteria, providing a high abundance of *Bacteroides* in these samples.

4.8.3. The Difference in Butyrate Levels Between Groups High in *Bacteroides* and High in *Bifidobacterium*

Butyrate is produced from many substrates. Some are lactate and acetate found in the pyruvate metabolism (figure 3.5C). However, butyrate levels were three times higher in the high in *Bacteroides* group compared to the high in *Bifidobacterium* group (figure 3.7), despite the two groups having almost equal acetate levels. This could reflect a difference in the abundance of, for example, lactate-utilizing species within the gut microbiota in the samples. There were particularly three samples that contributed to the high butyrate levels, and these were samples 2, 3, and 4 (figure 3.6). It is noteworthy that those three samples all were high in *Bacteroides* samples and that the correlation analysis had an, although weak, negative correlation between butyrate and *Bifidobacterium*. This could indicate that *Bifidobacterium* has a negative effect on butyrate production. But, the correlation analysis had ($p < 0.1$), and the correlation was not significant. An analysis with more samples would be necessary for a more accurate result, and the p-value should be lower than 0.05.

4.9. Technical Considerations and Future Research

4.9.1. Shotgun Analysis

The most abundant species of *Bacteroides* and *Bifidobacterium* found in the shotgun data (Figure 3.3) were used to validate the data. These were *Bacteroides fragilis* and *Bifidobacterium longum*. The KEGG pathway database provides metabolic pathways known for bacterial species, and the galactose degradation pathway, as well as the succinate pathway of *B.fragilis* and *B. longum,* were investigated to compare with the shotgun data. KEGG pathway database provided the same information as observed in both pathways of shotgun data (figure 3.5A and 3.5D). This offers higher credibility to the shotgun data.

For a broader understanding of mucin and HMO utilization, it would be interesting to investigate specific species of the two genera. This thesis did not focus on the specific species but the genus level and the potential metabolism present in each genus. Most likely, not all species of *Bacteroides* and *Bifidobacterium* that inhabit the infant gut possess all the pathways or GHs presented in this thesis. They may depend on cross-feeding strategies between themselves and other species, and knowing the contribution of each specie would provide for a better understanding of their metabolism in the future.

4.9.2. The Proteome Analysis

There was high concordance of the two parallels run in the protein extraction. In addition, the clusters with proteins annotated to *Bifidobacterium* had a higher abundance within samples from the high in *Bifidobacterium* group and vice versa (figure 3.4C). These observations made the results more trustworthy. However, the number of proteins varied between the samples, and barely any proteins were extracted from samples 3, 4, 8, and 9 compared to the other samples (Table 3.2). This could be due to repeated thawing and freezing of samples before use, and the protein extraction protocol is extensive in terms of potential missteps, which could have influenced the result. The analysis only provided proteins from inside the bacterial cells, and the number of extracellular proteins in the bacteria could impact the number of proteins that were extracted. Substantially fewer proteins were extracted from the *Bifidobacterium* genus, and it may have a higher quantity of extracellular proteins than *Bacteroides*. Additionally, the *Bifidobacterium* is a gram-positive bacterium in contrast to the gram-negative *Bacteroides* and could be more difficult to lyse. The sample size of 11 samples could also be a limitation regarding protein extraction, and more samples would provide for higher variation in the bacterial composition. This would increase the possibility of obtaining more proteins, including more unique proteins from the protein extraction.

A final limitation is the increased risk of false positives when mapping bacteria to the proteins discovered in the mass spectrometry because all observed proteins intended to map to the species in the shotgun database. The database was based on species from *Bifidobacterium* and *Bacteroides*, and proteins of other species that resembles the amino acid sequence of *Bifidobacterium* or *Bacteroides* proteins may annotate to them after all.

5. Conclusion

The shotgun analysis uncovered the full presence of the glycoside hydrolases necessary for HMO degradation for *Bacteroides* and *Bifidobacterium*, whereas many of the discovered GHs may also apply to mucin degradation. This indicates a common potential for *Bacteroides* and *Bifidobacterium* to utilize these substrates. The proteome analysis revealed a presence of fucosidase, which removes fucose from HMOs and mucins in both genera. The β-hexosaminidase was found in the proteome for *Bacteroides* but was lacking in *Bifidobacterium*. The β-hexosaminidase Lacto-N-biosidase, releasing lactose from the reducing end of HMOs, was observed in the proteomics for both genera. Further, to degrade lactose, β-galactosidases are required, and the protein was only found in the proteome of *Bifidobacterium*. The GLNBP was detected only in *Bifidobacterium*. However, *Bacteroides* may exploit N-acetyl-glucosaminidases instead of GLNBP. And finally, the sialidase was lacking in the proteomics of both genera. An extended sample set with more samples providing a wider bacterial composition of *Bacteroides* and *Bifidobacterium* may lead to the detection of more proteins relevant for HMO- and mucin degradation. The variation of enzymes present in *Bifidobacterium* likely reveals a better adaption towards the utilization of HMO than *Bacteroides*, and it seems like the two genera overlap in their utilization specialties. In addition to its mucin-degrading capabilities, there were indications for the ability of *Bacteroides* to utilize other substrates with sulfatases, like GAGs, which are also highly present in human breast milk. The metabolic pathways and potential competition for common resources between *Bacteroides* and *Bifidobacterium* should be further investigated.

6. References

Almagro-Moreno, S. & Boyd, E. F. (2009). Insights into the evolution of sialic acid catabolism among bacteria. *BMC Evolutionary Biology*, 9 (1): 118. doi: 10.1186/1471-2148-9-118.

Avershina, E., Lundgard, K., Sekelja, M., Dotterud, C., Storro, O., Oien, T., Johnsen, R. & Rudi, K. (2016). Transition from infant- to adult-like gut microbiota. *Environ Microbiol*, 18 (7): 2226-36. doi: 10.1111/1462-2920.13248.

Backhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., et al. (2015). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe*, 17 (5): 690-703. doi: 10.1016/j.chom.2015.04.004.

Bakken, L. R., Bergaust, L., Liu, B. & Frostegard, A. (2012). Regulation of denitrification at the cellular level: a clue to the understanding of N2O emissions from soils. *Philos Trans R Soc Lond B Biol Sci*, 367 (1593): 1226-34. doi: 10.1098/rstb.2011.0321.

Bansil, R. & Turner, B. S. (2006). Mucin structure, aggregation, physiological functions and biomedical applications. *Current Opinion in Colloid & Interface Science*, 11 (2-3): 164-170. doi: 10.1016/j.cocis.2005.11.001.

Bell, A., Brunt, J., Crost, E., Vaux, L., Nepravishta, R., Owen, C. D., Latousakis, D., Xiao, A., Li, W., Chen, X., et al. (2019). Elucidation of a sialic acid metabolism pathway in mucus-foraging Ruminococcus gnavus unravels mechanisms of bacterial adaptation to the gut. *Nature Microbiology*, 4 (12): 2393-2404. doi: 10.1038/s41564-019-0590-7.

Bell, A. & Juge, N. (2021). Mucosal glycan degradation of the host by the gut microbiota. *Glycobiology*, 31 (6): 691-696. doi: 10.1093/glycob/cwaa097.

Bode, L. (2012). Human milk oligosaccharides: every baby needs a sugar mama. *Glycobiology*, 22 (9): 1147-62. doi: 10.1093/glycob/cws074.

Borewicz, K., Gu, F., Saccenti, E., Arts, I. C. W., Penders, J., Thijs, C., van Leeuwen, S. S., Lindner, C., Nauta, A., van Leusen, E., et al. (2019). Correlating Infant Faecal Microbiota Composition and Human Milk Oligosaccharide Consumption by Microbiota of One-Month Old Breastfed Infants. *Mol Nutr Food Res*: e1801214. doi: 10.1002/mnfr.201801214.

Bunesova, V., Lacroix, C. & Schwab, C. (2018). Mucin Cross-Feeding of Infant Bifidobacteria and Eubacterium hallii. *Microb Ecol*, 75 (1): 228-238. doi: 10.1007/s00248-017-1037-4.

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, 38 (12): 5825-5829. doi: 10.1093/molbev/msab293.

Carrow, H. C., Batachari, L. E. & Chu, H. (2020). Strain diversity in the microbiome: Lessons from Bacteroides fragilis. *PLOS Pathogens*, 16 (12): e1009056. doi: 10.1371/journal.ppat.1009056.

Collado, M. C., Rautava, S., Aakko, J., Isolauri, E. & Salminen, S. (2016). Human gut colonisation may be initiated in utero by distinct microbial communities in the placenta and amniotic fluid. *Sci Rep*, 6: 23129. doi: 10.1038/srep23129.

Dau, T., Bartolomucci, G. & Rappsilber, J. (2020). Proteomics Using Protease Alternatives to Trypsin Benefits from Sequential Digestion with Trypsin. *Anal Chem*, 92 (14): 9523-9527. doi: 10.1021/acs.analchem.0c00478.

De Bruyn, F., Beauprez, J., Maertens, J., Soetaert, W. & De Mey, M. (2013). Unraveling the Leloir pathway of Bifidobacterium bifidum: significance of the uridylyltransferases. *Appl Environ Microbiol*, 79 (22): 7028-35. doi: 10.1128/AEM.02460-13.

den Besten, G., van Eunen, K., Groen, A. K., Venema, K., Reijngoud, D. J. & Bakker, B. M. (2013). The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *J Lipid Res*, 54 (9): 2325-40. doi: 10.1194/jlr.R036012.

Domon, B. & Aebersold, R. (2006). Mass spectrometry and protein analysis. *Science*, 312. doi: 10.1126/science.1124619.

Drula, E., Garron, M.-L., Dogan, S., Lombard, V., Henrissat, B. & Terrapon, N. (2021). The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Research*, 50 (D1): D571-D577. doi: 10.1093/nar/gkab1045.

Eisenstein, M. (2020). The hunt for a healthy microbiome. *Springer Nature Limited*, 577: S6-S8 (2020). doi: https://doi.org/10.1038/d41586-020-00193-3.

Escobar-Zepeda, A., Vera-Ponce de Leon, A. & Sanchez-Flores, A. (2015). The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Front Genet*, 6: 348. doi: 10.3389/fgene.2015.00348.

Forster, S. C., Kumar, N., Anonye, B. O., Almeida, A., Viciani, E., Stares, M. D., Dunn, M., Mkandawire, T. T., Zhu, A., Shao, Y., et al. (2019). A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat Biotechnol*, 37 (2): 186-192. doi: 10.1038/s41587-018-0009-7.

Fujita, K., Oura, F., Nagamine, N., Katayama, T., Hiratake, J., Sakata, K., Kumagai, H. & Yamamoto, K. (2005). Identification and molecular cloning of a novel glycoside hydrolase family of core 1 type O-glycan-specific endo-alpha-N-acetylgalactosaminidase from Bifidobacterium longum. *J Biol Chem*, 280 (45): 37415-22. doi: 10.1074/jbc.M506874200.

Gotoh, A., Katoh, T., Sakanaka, M., Ling, Y., Yamada, C., Asakuma, S., Urashima, T., Tomabechi, Y., Katayama-Ikegami, A., Kurihara, S., et al. (2018). Sharing of human milk oligosaccharides degradants within bifidobacterial communities in faecal cultures supplemented with Bifidobacterium bifidum. *Sci Rep*, 8 (1): 13958. doi: 10.1038/s41598-018-32080-3.

Granger, C. L., Embleton, N. D., Palmer, J. M., Lamb, C. A., Berrington, J. E. & Stewart, C. J. (2021). Maternal breastmilk, infant gut microbiome and the impact on preterm infant health. *Acta Paediatr*, 110 (2): 450-457. doi: 10.1111/apa.15534.

Heather, J. M. & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107 (1): 1-8. doi: 10.1016/j.ygeno.2015.11.003.

Hoang, D. M., Levy, E. I. & Vandenplas, Y. (2021). The impact of Caesarean section on the infant gut microbiome. *Acta Paediatrica*, 110 (1): 60-67. doi: https://doi.org/10.1111/apa.15501.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, Lars J., et al. (2018). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47 (D1): D309-D314. doi: 10.1093/nar/gky1085.

Hur, K. Y. & Lee, M. S. (2015). Gut Microbiota and Metabolic Disorders. *Diabetes Metab J*, 39 (3): 198-203. doi: 10.4093/dmj.2015.39.3.198.

Illumina, Inc. (2017). An introduction to Next Generation Sequencing Technology. *Illumina*, Available at: https://www.illumina.com/documents/products/illumina_sequencing_introduction.pdf.

Ioannou, A., Knol, J. & Belzer, C. (2021). Microbial Glycoside Hydrolases in the First Year of Life: An Analysis Review on Their Presence and Importance in Infant Gut. *Front Microbiol*, 12: 631282. doi: 10.3389/fmicb.2021.631282.

Kanehisa, M. & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28 (1): 27-30. doi: 10.1093/nar/28.1.27.

Kanehisa, M., Sato, Y. & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol*, 428 (4): 726-731. doi: 10.1016/j.jmb.2015.11.006.

Kononova, S., Litvinova, E., Vakhitov, T., Skalinskaya, M. & Sitkin, S. (2021). Acceptive Immunity: The Role of Fucosylated Glycans in Human Host-Microbiome Interactions. *Int J Mol Sci*, 22 (8). doi: 10.3390/ijms22083854.

Kostopoulos, I., Elzinga, J., Ottman, N., Klievink, J. T., Blijenberg, B., Aalvink, S., Boeren, S., Mank, M., Knol, J., de Vos, W. M., et al. (2020). Akkermansia muciniphila uses human milk oligosaccharides to thrive in the early life conditions in vitro. *Scientific Reports*, 10 (1): 14330. doi: 10.1038/s41598-020-71113-8.

Koutsioulis, D., Landry, D. & Guthrie, E. P. (2008). Novel endo-α-N-acetylgalactosaminidases with broader substrate specificity. *Glycobiology*, 18 (10): 799-805. doi: 10.1093/glycob/cwn069.

Li, Z. & Chai, W. (2019). Mucin O-glycan microarrays. *Curr Opin Struct Biol*, 56: 187-197. doi: 10.1016/j.sbi.2019.03.032.

Lin, C. S., Chang, C. J., Lu, C. C., Martel, J., Ojcius, D. M., Ko, Y. F., Young, J. D. & Lai, H. C. (2014). Impact of the gut microbiota, prebiotics, and probiotics on human health and disease. *Biomed J*, 37 (5): 259-68. doi: 10.4103/2319-4170.138314.

Lodrup Carlsen, K. C., Rehbinder, E. M., Skjerven, H. O., Carlsen, M. H., Fatnes, T. A., Fugelli, P., Granum, B., Haugen, G., Hedlin, G., Jonassen, C. M., et al. (2018). Preventing Atopic Dermatitis and ALLergies in Children-the PreventADALL study. *Allergy*, 73 (10): 2063-2070. doi: 10.1111/all.13468.

Louis, P., Scott, K. P., Duncan, S. H. & Flint, H. J. (2007). Understanding the effects of diet on bacterial metabolism in the large intestine. *J Appl Microbiol*, 102 (5): 1197-208. doi: 10.1111/j.1365-2672.2007.03322.x.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S. & Shafee, T. (2017). Transcriptomics technologies. *PLOS Computational Biology*, 13 (5): e1005457. doi: 10.1371/journal.pcbi.1005457.

Luis, A. S., Jin, C., Pereira, G. V., Glowacki, R. W. P., Gugel, S. R., Singh, S., Byrne, D. P., Pudlo, N. A., London, J. A., Basle, A., et al. (2021). A single sulfatase is required to access colonic mucin by a gut bacterium. *Nature*, 598 (7880): 332-337. doi: 10.1038/s41586-021-03967-5.

Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., Greenhalgh, K., Jäger, C., Baginska, J., Wilmes, P., et al. (2017). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature Biotechnology*, 35 (1): 81-89. doi: 10.1038/nbt.3703.

Makino, H., Kushiro, A., Ishikawa, E., Kubota, H., Gawad, A., Sakai, T., Oishi, K., Martin, R., Ben-Amor, K., Knol, J., et al. (2013). Mother-to-Infant Transmission of Intestinal Bifidobacterial Strains Has an Impact on the Early Development of Vaginally Delivered Infant's Microbiota. *PLOS ONE*, 8 (11): e78331. doi: 10.1371/journal.pone.0078331.

Marcobal, A., Barboza, M., Sonnenburg, E. D., Pudlo, N., Martens, E. C., Desai, P., Lebrilla, C. B., Weimer, B. C., Mills, D. A., German, J. B., et al. (2011). Bacteroides in the infant gut consume milk oligosaccharides via mucus-utilization pathways. *Cell Host Microbe*, 10 (5): 507-14. doi: 10.1016/j.chom.2011.10.007.

Martin-Gallausiaux, C., Marinelli, L., Blottière, H. M., Larraufie, P. & Lapaque, N. (2021). SCFA: mechanisms and functional importance in the gut. *Proceedings of the Nutrition Society*, 80 (1): 37-49. doi: 10.1017/S0029665120006916.

Masi, A. C. & Stewart, C. J. (2022). Untangling human milk oligosaccharides and infant gut microbiome. *iScience*, 25 (1): 103542. doi: 10.1016/j.isci.2021.103542.

Matamoros, S., Gras-Leguen, C., Le Vacon, F., Potel, G. & de La Cochetiere, M. F. (2013). Development of intestinal microbiota in infants and its impact on health. *Trends Microbiol*, 21 (4): 167-73. doi: 10.1016/j.tim.2012.12.001.

Milani, C., Lugli, G. A., Duranti, S., Turroni, F., Bottacini, F., Mangifesta, M., Sanchez, B., Viappiani, A., Mancabelli, L., Taminiau, B., et al. (2014). Genomic encyclopedia of type strains of the genus Bifidobacterium. *Appl Environ Microbiol*, 80 (20): 6290-302. doi: 10.1128/aem.02308-14.

Milani, C., Turroni, F., Duranti, S., Lugli, G. A., Mancabelli, L., Ferrario, C., van Sinderen, D. & Ventura, M. (2016). Genomics of the Genus Bifidobacterium Reveals Species-Specific Adaptation to the Glycan-Rich Gut Environment. *Appl Environ Microbiol*, 82 (4): 980-991. doi: 10.1128/AEM.03500-15.

Milani, C., Duranti, S., Bottacini, F., Casey, E., Turroni, F., Mahony, J., Belzer, C., Delgado Palacio, S., Arboleya Montes, S., Mancabelli, L., et al. (2017). The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. *Microbiol Mol Biol Rev*, 81 (4). doi: 10.1128/mmbr.00036-17.

Miwa, M., Horimoto, T., Kiyohara, M., Katayama, T., Kitaoka, M., Ashida, H. & Yamamoto, K. (2010). Cooperation of β-galactosidase and β-N-acetylhexosaminidase from bifidobacteria in assimilation of human milk oligosaccharides with type 2 structure. *Glycobiology*, 20 (11): 1402-1409. doi: 10.1093/glycob/cwq101.

Mosca, F. & Gianni, M. L. (2017). Human milk: composition and health benefits. *Pediatr Med Chir*, 39 (2): 155. doi: 10.4081/pmc.2017.155.

Nilsen, M., Lokmic, A., Angell, I. L., Lodrup Carlsen, K. C., Carlsen, K. H., Haugen, G., Hedlin, G., Jonassen, C. M., Marsland, B. J., Nordlund, B., et al. (2021). Fecal Microbiota Nutrient Utilization Potential Suggests Mucins as Drivers for Initial Gut Colonization of Mother-Child-Shared Bacteria. *Appl Environ Microbiol*, 87 (6). doi: 10.1128/AEM.02201-20.

Nishimoto, M. & Kitaoka, M. (2007). Identification of N-acetylhexosamine 1-kinase in the complete lacto-N-biose I/galacto-N-biose metabolic pathway in Bifidobacterium longum. *Appl Environ Microbiol*, 73 (20): 6444-9. doi: 10.1128/AEM.01425-07.

Parte, A. C., Carbasse, J. S., 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): 5608-5612. doi: 10.1099/ijsem.0.004332.

Perez-Munoz, M. E., Arrieta, M. C., Ramer-Tait, A. E. & Walter, J. (2017). A critical assessment of the "sterile womb" and "in utero colonization" hypotheses: implications for research on the pioneer infant microbiome. *Microbiome*, 5 (1): 48. doi: 10.1186/s40168-017-0268-4.

Primec, M., Micetic-Turk, D. & Langerholc, T. (2017). Analysis of short-chain fatty acids in human feces: A scoping review. *Anal Biochem*, 526: 9-21. doi: 10.1016/j.ab.2017.03.007.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464 (7285): 59-65. doi: 10.1038/nature08821.

Quin, C., Vicaretti, S. D., Mohtarudin, N. A., Garner, A. M., Vollman, D. M., Gibson, D. L. & Zandberg, W. F. (2020). Influence of sulfonated and diet-derived human milk oligosaccharides on the infant microbiome and immune markers. *J Biol Chem*, 295 (12): 4035-4048. doi: 10.1074/jbc.RA119.011351.

Raimondi, S., Musmeci, E., Candeliere, F., Amaretti, A. & Rossi, M. (2021). Identification of mucin degraders of the human gut microbiota. *Sci Rep*, 11 (1): 11094. doi: 10.1038/s41598-021-90553-4.

Rautava, S., Luoto, R., Salminen, S. & Isolauri, E. (2012). Microbial contact during pregnancy, intestinal colonization and human disease. *Nat Rev Gastroenterol Hepatol*, 9 (10): 565-76. doi: 10.1038/nrgastro.2012.144.

Ravi, A., Avershina, E., Angell, I. L., Ludvigsen, J., Manohar, P., Padmanaban, S., Nachimuthu, R., Snipen, L. & Rudi, K. (2018). Comparison of reduced metagenome and 16S rRNA gene sequencing for determination of genetic diversity and mother-child overlap of the gut associated microbiota. *J Microbiol Methods*, 149: 44-52. doi: 10.1016/j.mimet.2018.02.016.

Rios-Covian, D., Gueimonde, M., Duncan, S. H., Flint, H. J. & de los Reyes-Gavilan, C. G. (2015). Enhanced butyrate formation by cross-feeding between Faecalibacterium prausnitzii and Bifidobacterium adolescentis. *FEMS Microbiol Lett*, 362 (21). doi: 10.1093/femsle/fnv176.

Rios-Covian, D., Salazar, N., Gueimonde, M. & de los Reyes-Gavilan, C. G. (2017). Shaping the Metabolism of Intestinal Bacteroides Population through Diet to Improve Human Health. *Frontiers in Microbiology*, 8. doi: 10.3389/fmicb.2017.00376.

Rutayisire, E., Huang, K., Liu, Y. & Tao, F. (2016). The mode of delivery affects the diversity and colonization pattern of the gut microbiota during the first year of infants' life: a systematic review. *BMC Gastroenterol*, 16 (1): 86. doi: 10.1186/s12876-016-0498-0.

Sabater, C., Ruiz, L. & Margolles, A. (2021). A Machine Learning Approach to Study Glycosidase Activities from Bifidobacterium. *Microorganisms*, 9 (5): 1034.

Sender, R., Fuchs, S. & Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *bioRxiv*: 036103. doi: 10.1101/036103.

Slatko, B. E., Gardner, A. F. & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Curr Protoc Mol Biol*, 122 (1): e59. doi: 10.1002/cpmb.59.

Tailford, L. E., Crost, E. H., Kavanaugh, D. & Juge, N. (2015). Mucin glycan foraging in the human gut microbiome. *Front Genet*, 6: 81. doi: 10.3389/fgene.2015.00081.

Tan, S. C. & Yiap, B. C. (2009). DNA, RNA, and protein extraction: the past and the present. *J Biomed Biotechnol*, 2009: 574398. doi: 10.1155/2009/574398.

Turroni, F., Bottacini, F., Foroni, E., Mulder, I., Kim, J. H., Zomer, A., Sánchez, B., Bidossi, A., Ferrarini, A., Giubellini, V., et al. (2010). Genome analysis of Bifidobacterium bifidum PRL2010 reveals metabolic pathways for host-derived glycan foraging. *Proc Natl Acad Sci U S A*, 107 (45): 19514-9. doi: 10.1073/pnas.1011100107.

Turroni, F., Peano, C., Pass, D. A., Foroni, E., Severgnini, M., Claesson, M. J., Kerr, C., Hourihane, J., Murray, D., Fuligni, F., et al. (2012). Diversity of Bifidobacteria within the Infant Gut Microbiota. *PLOS ONE*, 7 (5): e36957. doi: 10.1371/journal.pone.0036957.

Turroni, F., Milani, C., Duranti, S., Mahony, J., van Sinderen, D. & Ventura, M. (2018). Glycan Utilization and Cross-Feeding Activities by Bifidobacteria. *Trends Microbiol*, 26 (4): 339-350. doi: 10.1016/j.tim.2017.10.001.

Tyers, M. & Mann, M. (2003). From genomics to proteomics. *Nature*, 422. doi: 10.1038/nature01510.

Vailati-Riboni, M., Palombo, V. & Loor, J. J. (2017). What are omics sciences? . In B., A. (ed.) *Periparturient Diseases of Dairy Cows*: Springer, Cham.

Ventura, M., Turroni, F., Lugli, G. A. & van Sinderen, D. (2014). Bifidobacteria and humans: our special friends, from ecological to genomics perspectives. *J Sci Food Agric*, 94 (2): 163-8. doi: 10.1002/jsfa.6356.

Vighi, G., Marcucci, F., Sensi, L., Di Cara, G. & Frati, F. (2008). Allergy and the gastrointestinal system. *Clin Exp Immunol*, 153 Suppl 1 (Suppl 1): 3-6. doi: 10.1111/j.1365-2249.2008.03713.x.

Wang, M., Zhao, Z., Zhao, A., Zhang, J., Wu, W., Ren, Z., Wang, P. & Zhang, Y. (2020). Neutral Human Milk Oligosaccharides Are Associated with Multiple Fixed and Modifiable Maternal and Infant Characteristics. *Nutrients*, 12 (3). doi: 10.3390/nu12030826.

Wexler, H. M. (2014). The Genus Bacteroides. In Rosenberg, E., DeLong, E. F., Lory, S., Stackebrandt, E. & Thompson, F. (eds) *The Prokaryotes: Other Major Lineages of Bacteria and The Archaea*, pp. 459-484. Berlin, Heidelberg: Springer Berlin Heidelberg.

Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. (2009). Universal sample preparation method for proteome analysis. *Nat Methods*, 6 (5): 359-62. doi: 10.1038/nmeth.1322.

Yu, Z. T., Chen, C. & Newburg, D. S. (2013). Utilization of major fucosylated and sialylated human milk oligosaccharides by isolated human gut microbes. *Glycobiology*, 23 (11): 1281-92. doi: 10.1093/glycob/cwt065.

Zafar, H. & Saier, M. H. (2021). Gut Bacteroides species in health and disease. *Gut Microbes*, 13 (1): 1848158. doi: 10.1080/19490976.2020.1848158.

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Busk, P. K., Xu, Y. & Yin, Y. (2018). dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, 46 (W1): W95-W101. doi: 10.1093/nar/gky418.

Appendix A1 – Tables with Sample Information and Measurements

Table 1. Relative abundance (0-1) of Bacteroides and Bifidobacterium in samples 1-11. These are numbers based on 16S rRNA sequencing. The table also illustrates the placement of samples for indexing later.

Placement	Sample#	*Bacteroides*	*Bifidobacterium*
A1	1	0,4110	0,00037
B1	2	0,5520	0,00072
C1	3	0,2252	0,001737
D1	4	0,3049	0,0461
E1	5	0,2485	0,1132
F1	6	0,4971	0,1481
A2	7	0,05096	0,3041
B2	8	0,00043	0,3676
C2	9	0,2531	0,4964
D2	10	0,00265	0,6038
E2	11	0,12050	0,6892
G1	12	Positive Control: *Bifidobacterium breve*	
F2	13	Neg. Control DNA- Extraction	
G2	14	Neg. Control PCR water	

Table 2. Index adapters for shotgun sequencing. A combination of i5 and i7 adapters were used on samples 1-14.

Index adapters	H503	H505	H506	H517
H705	A1	B1	C1	D1
H706	E1	F1	G1	A2
H707	B2	C2	D2	E2
H710	F2	G2	-	-

Table 3. Overview of grams feces added to falcon tubes of samples 1-11. The samples were different and certain were difficult to weigh out, including some of the samples not having enough feces left for 2x 0.2g.

Highest % bacteria	Sample #	New sample name and measured grams		Replicate name and measured grams	
Bacteroides	1	1.1a	0.195	1.1b	0.191
Bacteroides	2	1.2a	0.202	1.2b	0.280
Bacteroides	3	1.3a	0.190	1.3b	0.202
Bacteroides	4	1.4a	0.196	1.4b	0.214
Bacteroides	5	1.5a	0.211	1.5b	0.181
Bacteroides	6	1.6a	0.236	1.6b	0.206
Bifidobacterium	7	2.1a	0.181	2.1b	0.131
Bifidobacterium	8	2.2a	0.204	2.2b	0.127
Bifidobacterium	9	2.3a	0.190	2.3b	0.210
Bifidobacterium	10	2.4a	0.222	2.4b	0.184
Bifidobacterium	11	2.5a	0.191	2.5b	0.205

Table 4. Protein concentrations. The measured concentration of protein samples, using Pierce BCA Protein Assay Kit. The dilution factor was 5 and must be multiplied by the concentration given by the bio photometer. The desired concentration was 2.05 µg/µl.

Sample	**Absorbance (562 nm)**	**Concentration (µg/ml) * dilution factor**	**Concentration (µg/µl)**
1.1a	0.231	275	0.275
1.1b	0.339	430	0.430
1.2a	0.567	750	0.750
1.2b	0.286	355	0.355
1.3a	0.354	450	0.450
1.3b	0.265	325	0.325
1.4a	0.158	170	0.170
1.4b	0.164	180	0.180
1.5a	0.170	190	0.190
1.5b	0.173	190	0.190
1.6a	0.143	150	0.150
1.6b	0.319	400	0.400
2.1a	0.631	840	0.840
2.1b	0.194	220	0.220

2.2a	0.236	280	0.280
2.2b	0.235	280	0.280
2.3a	0.335	420	0.420
2.3b	0.192	220	0.220
2.4a	0.604	805	0.805
2.4b	0.654	875	0.875
2.5a	0.140	145	0.145
2.5b	0.169	185	0.185

Table 5. **NanoDrop measurement.** All 11 samples with replicates were measured on NanoDrop.

Sample	Mg/mL	A280	A260/A280
1.1a	0.023	0.02	1.75
1.1b	0.017	0.02	1.63
1.2a	0.040	0.04	1.68
1.2b	0.025	0.03	1.82
1.3a	0.025	0.03	1.73
1.3b	0.025	0.03	1.31
1.4a	0.024	0.02	1.96
1.4b	0.031	0.03	1.42
1.5a	0.032	0.03	1.84
1.5b	0.020	0.02	1.94
1.6a	0.008	0.01	2.90
1.6b	0.008	0.01	2.94
2.1a	0.017	0.02	1.69
2.1b	0	0	0
2.2a	0.004	0	3.80
2.2b	0.022	0.02	2.01
2.3a	0.004	0	4.85
2.3b	0.007	0.01	2.42
2.4a	0.01	0.01	2.73
2.4b	0.054	0.05	1.70
2.5a	0.020	0.02	2.06
2.5b	0.026	0.03	1.64

Appendix A2 – Figures

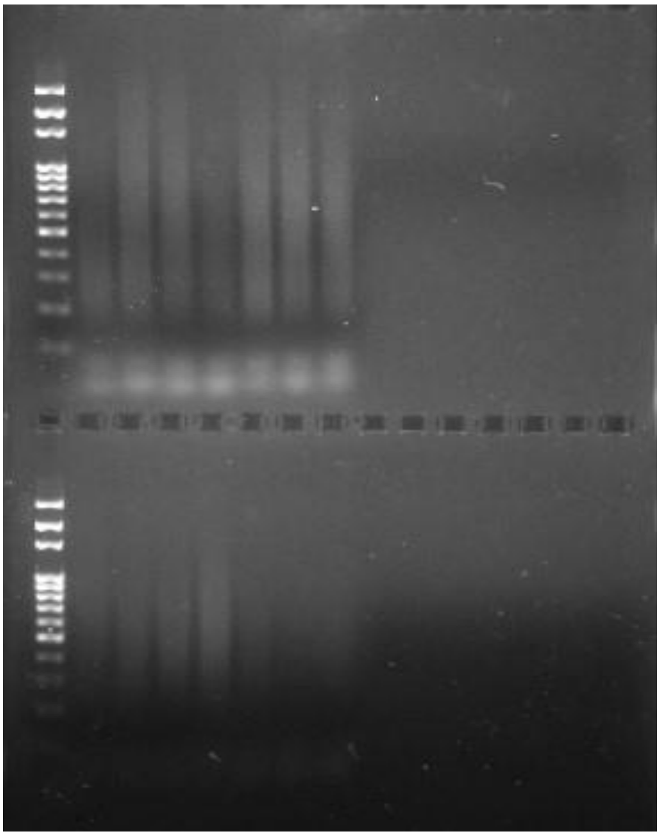

Figure 1. Gel picture of all samples after amplification for shotgun sequencing. Upper lanes: 100bp ladder, sample 1-6 + positive control. Lower lanes: 100bp ladder, sample 7-11 + two negative controls. The smear indicates fragmented DNA, and fragment sizes ranges from ~150 bp to 3000bp. The bottom of the wells may be debris from buffers or other contaminants.

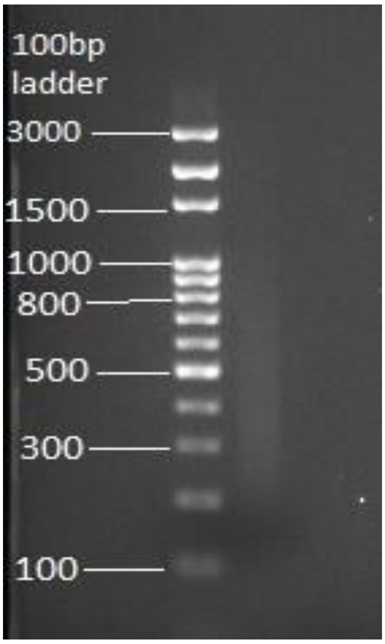

Figure 2. Pooled library gel picture before shotgun sequencing. The smear is caused by the different fragment sizes of the intended size from 200-1000bp. Illumina sequence fragments up to around 500bp and will not be able to sequence overlapping sequences that are bigger than that.

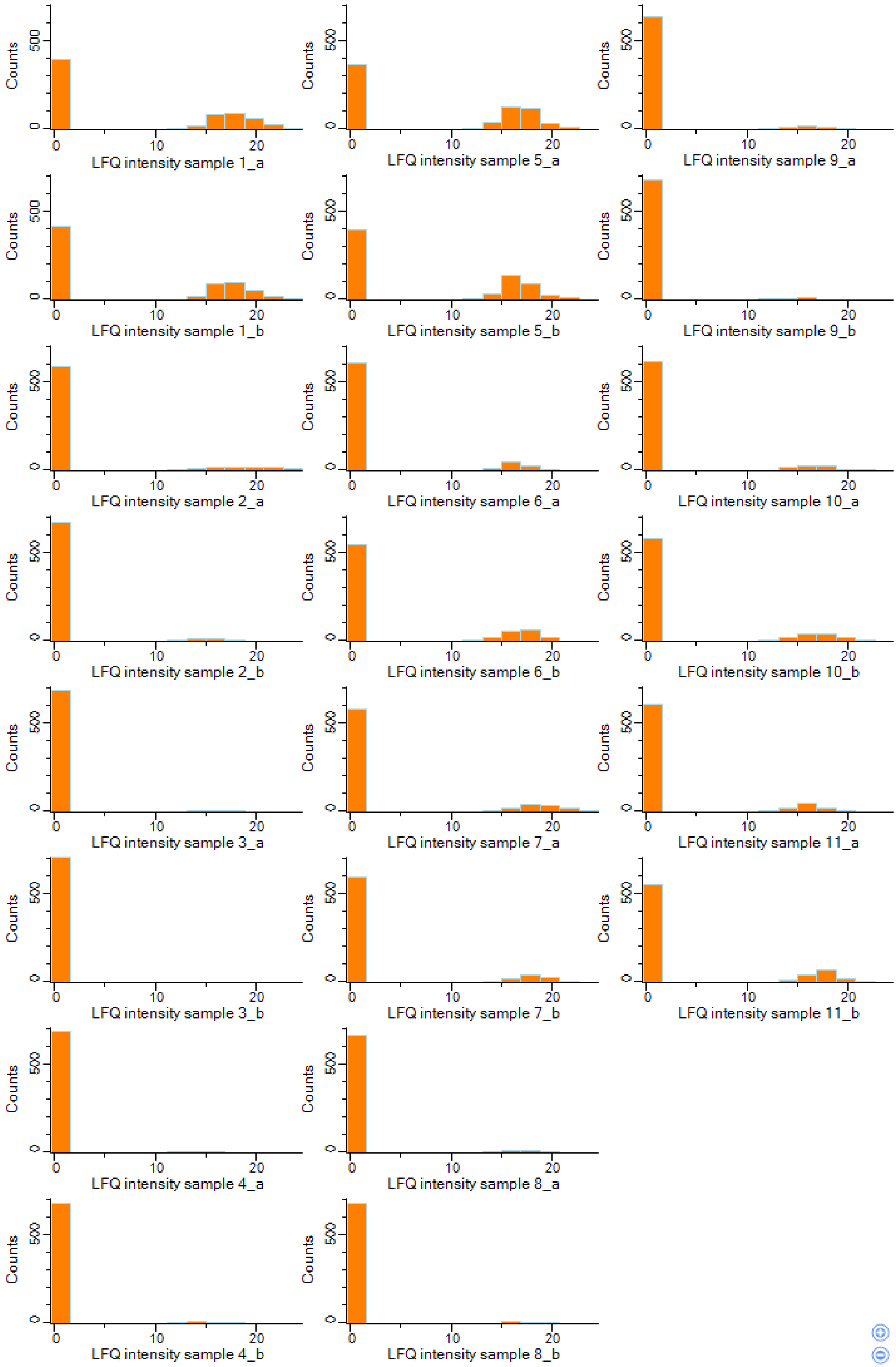

Figure 3. Histograms with LFQ intensity in samples 1-11 with replicates (a/b). The 0-block is replacing the NaN missing values. The figure represents the number of unique proteins that are present in the samples and illustrates why some samples were removed. The figure was made in Perseus version 1.6.6.0.

Appendix B – Protocols

Protocol B1.

GC analysis of short chain fatty acids in fecal samples

Instrument: Trace 1310 with autosampler (ThermoFisher Scientific)

Injector:

Mode: split

Temperature: 250 °C

Carrier gas: Helium

Column flow: 2.5 ml/min

Split flow: 200 ml/min

Purge flow: 3 ml/min

Injection volume: 0.2 µl

Liner: 4mm x 6.3mm x 78.5mm (Catalog# 23311.5, Restek)

Syringe: 10 µl syr FN 50 mm C, Ga 23, cone tip (catalog# 365D3741, ThermoFisher Scientific)

Column:

Stabilwax DA 30m, 0.25 mm ID, 0.25 µm (Restek)

Temperature program: 90 °C to 150 °C (6 min), 150 °C to 245 °C (1.9 min)

Time per sample: 14.9 min

Detector:

Type: FID

Temperature: 275 °C

Hydrogen: 30 ml/min

Air: 300 ml/min

Makeup gas: 30 ml/min

Protocol B2.

Mass Spectrometry - TimsTOF Aurora

The peptide samples were analysed by coupling a nano UPLC (nanoElute, Bruker) to a trapped ion mobility spectrometry/quadrupole time of flight mass spectrometer (timsTOF Pro, Bruker). The peptides were separated by an Aurora C18 reverse-phase (1.6 µm, 120Å) 25 cm X 75 µm analytical column with an integrated emitter (IonOpticks, Melbourne, Australia). The temperature of the column was kept at 50°C using the integrated oven. Equilibration of the column was performed before the samples were loaded (equilibration pressure 800 bar). The flow rate was set to 300 nl/min and the samples were separated using a solvent gradient from 2% to 25 % solvent B over 70 minutes, and to 37 % over 9 minutes. The solvent composition was then increased to 95 % solvent B over 10 min and maintained at that level for an additional 10 min. In total, a run time of 99 min was used for the separation of the peptides. Solvent A is 0.1 % (v/v) formic acid in milliQ water, while solvent B is 0.1 % (v/v) formic acid in LC-MS grade acetonitrile.

The timsTOF Pro was run in positive ion data-dependent acquisition PASEF mode with the control software Compass Hystar version 5.1.8.1 and timsControl version 1.1.19 68. The acquisition mass range was set to 100 – 1700 m/z. The TIMS settings were: 1/K0 Start 0.85 V·s/cm2 and 1/K0 End 1.4 V·s/cm2, ramp time 100 ms, ramp rate 9.42 Hz, and duty cycle 100 %. The capillary voltage was set at 1400 V, dry gas at 3.0 l/min, and dry temp at 180 ℃. The MS/MS settings were the following: number of PASEF ramps 10, total cycle time 0.53 sec, charge range 0-5, scheduling target intensity 20000, intensity threshold 2500, active exclusion release after 0.4 min, and CID collision energy ranging from 27-45 eV.

Appendix C – R Markdown-files

C1 – Annotation of Bacterial Taxonomy with the HumGut Database

Downloaded the HumGut table with metadata for each HumGut genome from [link] http://arken.nmbu.no/~larssn/humgut/ Also removed unnecessary columns.

```
library(readr)

## Warning: package 'readr' was built under R version 4.1.2

HumGut.tsv <- read.table(file = "HumGut.tsv", sep = "\t", header = TRUE)
HumGut.tsv <- HumGut.tsv[,-2:-8]
HumGut.tsv <- HumGut.tsv[,-3:-15]
```

Loaded the "Reads-shotgun file"/"reads_file" containing all sequences from Shotgun sequencing and removing values in parentheses etcetera.

```
reads_file <- "reads.txt"
new_krk.tbl <- read_delim(reads_file, delim = "\t",
                          col_names = c("C/U","Seq.ID","Tax.ID","bp.length"
,"LCA"),
                          trim_ws = T)

## Rows: 11739 Columns: 5
## -- Column specification ------------------------------------------------
--------
## Delimiter: "\t"
## chr (4): C/U, Seq.ID, Tax.ID, LCA
## dbl (1): bp.length
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.

new_krk.tbl$Tax.ID <- gsub("\\s*\\([^\\)]+\\)", "", new_krk.tbl$Tax.ID)
new_krk.tbl <- new_krk.tbl[,-5]
new_krk.tbl$OrganismName <- "NA"
colnames(HumGut.tsv) = c("Tax.ID", "OrganismName")
```

Want to match HumGut-IDs from the HumGut-database with the HumGut-IDs in "new_krk.tbl", and then copy these into the new column "Organismname".

```
library(stringr)

## Warning: package 'stringr' was built under R version 4.1.2

all_humgut <- str_detect(new_krk.tbl$Tax.ID, pattern = "HumGut")
all_to_get_annotations <- which(all_humgut == TRUE)
for (i in all_to_get_annotations) {
  matcher <- which(new_krk.tbl$Tax.ID[i] == HumGut.tsv$Tax.ID)
  OrganismeNavnet <- HumGut.tsv$OrganismName[matcher]
  new_krk.tbl$OrganismName[i] <- OrganismeNavnet
}
```

Wants to check which cells that have no ID, and that the taxonomy is correct.

```
no_humgut <- str_detect(new_krk.tbl$OrganismName, pattern = "NA")
indeks <- which(no_humgut == TRUE)

for (i in indeks) {
  new_krk.tbl$OrganismName[i] <- new_krk.tbl$Tax.ID[i]
}

dobbeltsjekk <- table(new_krk.tbl$Tax.ID[all_to_get_annotations], new_krk.t
bl$OrganismName[all_to_get_annotations])
dobbeltsjekk <- as.data.frame(dobbeltsjekk)
# View(dobbeltsjekk)
# View(HumGut.tsv)
```

Wants to check out the unique nodes:

```
head(unique(new_krk.tbl$Seq.ID))

## [1] "NODE_18_length_270080_cov_476.099574"
## [2] "NODE_41_length_194961_cov_451.282179"
## [3] "NODE_50_length_184487_cov_444.522800"
## [4] "NODE_63_length_153841_cov_421.194703"
## [5] "NODE_69_length_147463_cov_438.804498"
## [6] "NODE_77_length_143847_cov_478.664781"
```

The annotation for bacterial taxonomy is done.

C2 – Get *Bifidobacterium* and *Bacteroides* Proteins from all Faa/Fasta-Files

First, all nodes belonging to the same BIN had to be checked if they belonged to the same species. To do this, all fasta-files from Drep had to be uploaded. A new data frame was made.

```
library(tidyverse)

lines <- list.files(".", pattern = "fa")

Table_nodebin <- matrix(nrow = 1, ncol = 2)
colnames(Table_nodebin) <- c('Bin', 'Nodes')
Table_nodebin <- tbl_df(Table_nodebin)
```

Checked which lines had nodes or not and retrieve them.

```
for (i in 1:length(lines)) {
  lines_Read <- readLines(lines[i])
  logicals <- str_detect(lines_Read, pattern = ">")
  idx <- which(logicals) #Tells us where

  Pre_Node_table <- matrix(data = NA, nrow = length(idx), ncol = 2)
  colnames(Pre_Node_table) <- c('Bin', 'Nodes')
  Pre_Node_table <- tbl_df(Pre_Node_table)
  Pre_Node_table$Bin <- "NA"
  Pre_Node_table$Nodes <- "NA"

  Pre_Node_table$Bin <- lines[i]

  Pre_Node_table$Nodes <-lines_Read[idx]
  Table_nodebin <- rbind(Pre_Node_table, Table_nodebin)
}
```

Remove all “ >” and then retrieve all species from new_krk.tbl to each node in Table_nodebin

```
Table_nodebin$Bacteria <- "NA"
Table_nodebin$Nodes <- str_remove_all(Table_nodebin$Nodes, pattern = ">")
rm_last_row <- nrow(Table_nodebin)
Table_nodebin <- Table_nodebin[-rm_last_row,]

for (i in 1:nrow(Table_nodebin)) {
  bacteria_inn <- which(new_krk.tbl$Seq.ID == Table_nodebin$Nodes[i])
  Table_nodebin$Bacteria[i] <- new_krk.tbl$OrganismName[bacteria_inn]
}
```

The latter code crosschecks random subjects from new_krk.tbl in Table_nodebin, for example:

```
Table_nodebin[260,]

## # A tibble: 1 x 3
##   Bin               Nodes                              Bacteria
##   <chr>             <chr>                              <chr>
## 1 Sample9.007.fasta NODE_228_length_12040_cov_2.708148 Actinomyces sp. p
h3

which(new_krk.tbl$Seq.ID == Table_nodebin$Nodes[260])

## [1] 11024

new_krk.tbl[11024,]

## # A tibble: 1 x 5
##   `C/U` Seq.ID                             Tax.ID       bp.length Organi
smName
##   <chr> <chr>                              <chr>            <dbl> <chr>
## 1 C     NODE_228_length_12040_cov_2.708148 HumGut_27048     12040 Actino
myces s~
```

We can see that they have the same node name and organism name, which is good.

However, we only wanted genes from species that were Bacteroides and Bifidobacterium, and then bind them into one table.

```
Idx_Bacteroides <- which(str_detect(Table_nodebin$Bacteria, pattern = "Bact
eroides*"))
```

```
Bacteroides_NodeBin <- Table_nodebin[Idx_Bacteroides,]

Idx_Parabacteroides <- which(str_detect(Table_nodebin$Bacteria, pattern = "
Parabacteroides*"))
Parabacteroides_NodeBin <- Table_nodebin[Idx_Parabacteroides,]

Idx_Bacteroidales <- which(str_detect(Table_nodebin$Bacteria, pattern = "Ba
cteroidales*"))
Bacteroidales_NodeBin <- Table_nodebin[Idx_Bacteroidales,]

Idx_Phocaeicola <- which(str_detect(Table_nodebin$Bacteria, pattern = "Phoc
aeicola*"))
Phocaeicola_NodeBin <- Table_nodebin[Idx_Phocaeicola,]

Idx_Bifido <- which(str_detect(Table_nodebin$Bacteria, pattern = "Bifido*")
)
Bifido_NodeBin <- Table_nodebin[Idx_Bifido,]
#Binding all tables together into one table
New_table <- rbind(Bifido_NodeBin, Bacteroides_NodeBin, Parabacteroides_Nod
eBin, Phocaeicola_NodeBin, Bacteroidales_NodeBin)

saveRDS(New_table, file = "Bacteroides_Bifido_tabell")
```

Had to check the bins from this table. These were the only ones that were used.

```
unike_bins <- unique(New_table$Bin)
```

A total of 42 bins have genes coming from either Bacteroides or Bifidobacterium species.

Had to check all nodes and which BINs they were from. 2975 nodes.

```
Ferdig_testbin <- NULL
for (i in 1:length(unike_bins)) {
  TestBin <- which(New_table$Bin == unike_bins[i])
  Ferdig_testbin <- append(Ferdig_testbin, TestBin)
  Testbin2 <- New_table$Bin[Ferdig_testbin]
  TestNode <- New_table$Nodes[Ferdig_testbin]
}
```

Making sure that it was correct, by looking at “testnode 28”

```
New_table$Nodes[Ferdig_testbin[28]]

## [1] "NODE_17_length_199593_cov_215.343654"

TestNode[28]

## [1] "NODE_17_length_199593_cov_215.343654"

unique(Testbin2)

##  [1] "Sample9.11.fa"      "Sample9.007.fasta"  "Sample9.003.fasta"
##  [4] "Sample9.001.fasta"  "Sample7.001.fasta"  "Sample6.020.fasta"
##  [7] "Sample6.019.fasta"  "Sample6.016.fasta"  "Sample6.014.fasta"
## [10] "Sample6.009.fasta"  "Sample6.008.fasta"  "Sample6.004.fasta"
## [13] "Sample5.4.fa"       "Sample5.27.fa"      "Sample5.15.fa"
## [16] "Sample4.21.fa"      "Sample4.2.fa"       "Sample4.19.fa"
## [19] "Sample4.005.fasta"  "Sample10.015.fasta" "Sample10.014.fasta"
## [22] "Sample10.011.fasta" "Sample10.010.fasta" "Sample10.009.fasta"
## [25] "Sample10.007.fasta" "Sample10.005.fasta" "Sample10.002.fasta"
## [28] "Sample1.013.fasta"  "Sample5.23.fa"      "Sample5.20.fa"
## [31] "Sample5.14.fa"      "Sample5.12.fa"      "Sample5.10.fa"
## [34] "Sample4.6.fa"       "Sample4.3.fa"       "Sample4.017.fasta"
## [37] "Sample4.006.fasta"  "Sample2.15.fa"      "Sample1.11.fa"
## [40] "Sample1.009.fasta"  "Sample1.001.fasta"  "Sample4.18.fa"
```

All unique bins were put in its own file “unike_bins” for read_lines later. To look at the proteins, faa-files must be used, as fasta-files are nucleotide-sequences.

```r
path_til_protein <- "Oda_shotgun/Drep/Data/prodigal/unike_bins/"
protein_lines <- list.files(path = path_til_protein, pattern = "fa")

protein_lines2 <- str_remove(protein_lines, pattern = ".faa") ##Making a te
mporary vector just to match files
```

Searched for all lines that matched with the unique bins.

```r
Alle_protlines_read <- NULL
for (i in 1:length(protein_lines2)) {
  Riktig_Bin <- which(protein_lines2 == unike_bins[i])
  denne_filen <- protein_lines[Riktig_Bin]
  protlines_Read <- readLines(paste0(path_til_protein, denne_filen))
  Alle_protlines_read <- append(Alle_protlines_read, protlines_Read)
}
```

The following code took about 30-45 minutes. The loop finds the lines in protlines_read that matches with test nodes, which are the nodes with Bifidobacterium and Bacteroides.

```r
Ferdig_variabel <- NULL
for (i in 1:length(TestNode)) {
  test_Variable <- which(str_detect(Alle_protlines_read, paste0(pattern = T
estNode[i], "*")) == TRUE)
  Ferdig_variabel <- append(Ferdig_variabel,test_Variable)
}
```

The following code shows how many lines were withdrawn from 42 bins and 2975 nodes

```r
length_table <- length(Alle_protlines_read[Ferdig_variabel])
print(paste("The table has ", length_table, " lines"))
```

Found lines with nodes and wanted to see where they were.

```r
prot_logicals <- str_detect(Alle_protlines_read, pattern = ">")
prot_idx <- which(prot_logicals)
#print(prot_idx)
```

Withdrew the sequences from 42 bins and 2975 nodes. This code took about 20-25 minutes to run.

```
Alle_sekvensene_vivilha <- NULL
for (i in 1:length(Ferdig_variabel)) {
  inni_protidx <- min(which(prot_idx > Ferdig_variabel[i]))
  sekvensene_vivilha <- Alle_protlines_read[c(Ferdig_variabel[i]:(prot_idx[
inni_protidx]-1))]
  Alle_sekvensene_vivilha <- append(Alle_sekvensene_vivilha,sekvensene_vivi
lha)
  #print(i)
}
```

C3 – Create a FASTA File for eggNOG Mapper Annotation

This FASTA file should contain the aminoacid sequences of all nodes that were from Bacteroides and Bifidobacterium species and will be run through eggNOG mapper for annotation, as well as dbCAN.

```
new_line <- ">"
Final_Alle_sekvensene_vivilha <- c(Alle_sekvensene_vivilha, new_line)


krok_munn <- which(str_detect(Final_Alle_sekvensene_vivilha, pattern = ">")
)
sink(file = "Skal_i_eggNOG.faa") ##Lager en fil som limer inn alt som egent
lig skal printes i konsollen (cat-funksjonen)
for (i in 1:length(krok_munn)) {
  protein_hoy <- min(which(krok_munn > krok_munn[i]))
  selve_linjen_hoy <- krok_munn[protein_hoy]-1
  selve_linjen_lav <- krok_munn[i]+1
  hele_prot_sekvensen <- Final_Alle_sekvensene_vivilha[selve_linjen_lav:sel
ve_linjen_hoy]
  Hente_ut_node <- Alle_sekvensene_vivilha[krok_munn[i]]
  #hele_prot_sekvensen <- cat(hele_prot_sekvensen, sep = "") #SlÃ�Â¥r samme
n linjene
  cat(Hente_ut_node)
  cat(sep = "\n")
  cat(hele_prot_sekvensen, sep = "")
  cat(sep = "\n")
}

sink() #Stenger filen: med annenhver node og proteinsekvensen
## revert output back to the console -- only then access the file!
closeAllConnections()
```

C4 – Skal_i_eggNOG.faa - File Annotated in eggNOG Mapper

```
library(readxl)

## Warning: package 'readxl' was built under R version 4.1.2

library(stringr)
eggNOG_results <- read_xlsx("EggNOG mapper results.xlsx")
colnames(eggNOG_results) <- eggNOG_results[5,]
eggNOG_results <- eggNOG_results[-c(1:5,70252:70254),] #Fjerner de to først
e kolonnene og de tre siste (NA)
colnames(eggNOG_results) <- str_replace(colnames(eggNOG_results), pattern =
"#", replacement = "")

  enkel_eggNOG_res <- eggNOG_results
  enkel_eggNOG_res <- enkel_eggNOG_res[,-c(2,5,6,14:18,20)]
```

Removed nodes with e-values higher than 1e-10, which were approximately half of the proteins.

```
lav_evalue <- which(enkel_eggNOG_res$evalue < 1e-10) #Wants these
head(enkel_eggNOG_res$evalue[lav_evalue])

## [1] "0.0"       "0.0"       "0.0"       "1.23e-191" "1.25e-287" "1.87e-4
0"

Høy_evalue <- which(enkel_eggNOG_res$evalue > 1e-10) #Removing these
head(enkel_eggNOG_res$evalue[Høy_evalue])

## [1] "6.28e-221" "4.83e-257" "2.11e-220" "8.18e-267" "8.48e-223" "2.46e-2
88"

enkel_eggNOG_res <- enkel_eggNOG_res[-c(Høy_evalue),] #Have 34 025 genes le
ft.
```

Wanted the taxonomy from the shotgun data (HumGut) in this table. Had to match node-names in the two tables:

```
Bacteroies_Bifido_tabell <- readRDS("Bacteroides_Bifido_tabell")
enkel_eggNOG_res$ShotgunTaks <- "NA"
enkel_eggNOG_res$ShotgunNode <- "NA"

One_digit <- which(str_detect(enkel_eggNOG_res$query, pattern = "NODE_[:dig
it:]_length_[:digit:][:digit:]"))
En_skal_inn <- str_extract(enkel_eggNOG_res$query[One_digit], pattern = "NO
DE_[:digit:]_length_[:digit:][:digit:]")
enkel_eggNOG_res$ShotgunNode[One_digit] <- En_skal_inn

Two_digits <- which(str_detect(enkel_eggNOG_res$query, pattern = "NODE_[:di
git:][:digit:]_length_[:digit:][:digit:]"))
To_skal_inn <- str_extract(enkel_eggNOG_res$query[Two_digits], pattern = "N
ODE_[:digit:][:digit:]_length_[:digit:][:digit:]")
enkel_eggNOG_res$ShotgunNode[Two_digits] <- To_skal_inn
```

```
Three_digits <- which(str_detect(enkel_eggNOG_res$query, pattern = "NODE_[:
digit:][:digit:][:digit:]_length_[:digit:][:digit:]"))
Tre_skal_inn <- str_extract(enkel_eggNOG_res$query[Three_digits], pattern =
"NODE_[:digit:][:digit:][:digit:]_length_[:digit:][:digit:]")
enkel_eggNOG_res$ShotgunNode[Three_digits] <- Tre_skal_inn

Four_digits <- which(str_detect(enkel_eggNOG_res$query, pattern = "NODE_[:d
igit:][:digit:][:digit:][:digit:]_length_[:digit:][:digit:]"))
Fire_skal_inn <- str_extract(enkel_eggNOG_res$query[Four_digits], pattern =
"NODE_[:digit:][:digit:][:digit:][:digit:]_length_[:digit:][:digit:]")
enkel_eggNOG_res$ShotgunNode[Four_digits] <- Fire_skal_inn

Five_digits <- which(str_detect(enkel_eggNOG_res$query, pattern = "NODE_[:d
igit:][:digit:][:digit:][:digit:][:digit:]_length_[:digit:][:digit:]"))
Fem_skal_inn <- str_extract(enkel_eggNOG_res$query[Five_digits], pattern =
"NODE_[:digit:][:digit:][:digit:][:digit:][:digit:]_length_[:digit:][:digit
:]")
enkel_eggNOG_res$ShotgunNode[Five_digits] <- Fem_skal_inn

#Må gjøre det samme med denne tabellen for å matche
Bacteroies_Bifido_tabell$ShotgunNode <- "NA"

One_digit_shot <- which(str_detect(Bacteroies_Bifido_tabell$Nodes, pattern
= "NODE_[:digit:]_length_[:digit:][:digit:]"))
En_skal_inn_shot <- str_extract(Bacteroies_Bifido_tabell$Nodes[One_digit_sh
ot], pattern = "NODE_[:digit:]_length_[:digit:][:digit:]")
Bacteroies_Bifido_tabell$ShotgunNode[One_digit_shot] <- En_skal_inn_shot

Two_digit_shot <- which(str_detect(Bacteroies_Bifido_tabell$Nodes, pattern
= "NODE_[:digit:][:digit:]_length_[:digit:][:digit:]"))
To_skal_inn_shot <- str_extract(Bacteroies_Bifido_tabell$Nodes[Two_digit_sh
ot], pattern = "NODE_[:digit:][:digit:]_length_[:digit:][:digit:]")
Bacteroies_Bifido_tabell$ShotgunNode[Two_digit_shot] <- To_skal_inn_shot

Three_digit_shot <- which(str_detect(Bacteroies_Bifido_tabell$Nodes, patter
n = "NODE_[:digit:][:digit:][:digit:]_length_[:digit:][:digit:]"))
Tre_skal_inn_shot <- str_extract(Bacteroies_Bifido_tabell$Nodes[Three_digit
_shot], pattern = "NODE_[:digit:][:digit:][:digit:]_length_[:digit:][:digit
:]")
Bacteroies_Bifido_tabell$ShotgunNode[Three_digit_shot] <- Tre_skal_inn_shot

Four_digit_shot <- which(str_detect(Bacteroies_Bifido_tabell$Nodes, pattern
= "NODE_[:digit:][:digit:][:digit:][:digit:]_length_[:digit:][:digit:]"))
Fire_skal_inn_shot <- str_extract(Bacteroies_Bifido_tabell$Nodes[Four_digit
_shot], pattern = "NODE_[:digit:][:digit:][:digit:][:digit:]_length_[:digit
:][:digit:]")
Bacteroies_Bifido_tabell$ShotgunNode[Four_digit_shot] <- Fire_skal_inn_shot

Five_digit_shot <- which(str_detect(Bacteroies_Bifido_tabell$Nodes, pattern
= "NODE_[:digit:][:digit:][:digit:][:digit:][:digit:]_length_[:digit:][:dig
it:]"))
Fem_skal_inn_shot <- str_extract(Bacteroies_Bifido_tabell$Nodes[Five_digit_
shot], pattern = "NODE_[:digit:][:digit:][:digit:][:digit:][:digit:]_length
_[:digit:][:digit:]")
Bacteroies_Bifido_tabell$ShotgunNode[Five_digit_shot] <- Fem_skal_inn_shot

## Finding the matching cells and put them into Bakterie_node_eggNOG$Shotgu
nTaks in a for-loop

for (i in 1:nrow(enkel_eggNOG_res)) {
  idx <- which(Bacteroies_Bifido_tabell$ShotgunNode == enkel_eggNOG_res$Sho
tgunNode[i])
  enkel_eggNOG_res$ShotgunTaks[i] <- Bacteroies_Bifido_tabell$Bacteria[idx]
}

enkel_eggNOG_res <- enkel_eggNOG_res[,-c(2,3)]

write.csv2(enkel_eggNOG_res, file = "EggNOG_gene_taxonomy.csv") #Saving the
file
```

To look at the COGs, the following codes were run. 87 different COG categories were observed which included the combinations of COGs, and 26 single COG categories exist. The goal was to find out which category had the most proteins, and all categories were checked. Examples:

```
unique(enkel_eggNOG_res$COG_category)

##  [1] "-"    "M"    "U"    "GM"   "G"    "O"    "P"    "S"    "J"    "F"
## [11] "H"    "K"    "E"    "KT"   "C"    "V"    "L"    "EGP"  "I"    "EG"
## [21] "EP"   "D"    "T"    "GK"   "EH"   "QT"   "GP"   "MV"   "LU"   "DG"
## [31] "JKL"  "KL"   "KLT"  "ET"   "EF"   "NU"   "LT"   "EK"   "DF"   "FG"
## [41] "FJ"   "LV"   "OP"   "DM"   "Q"    "IQ"   "IM"   "A"    "CE"   "EQ"
## [51] "CO"   "N"    "PT"   "MU"   "FK"   "EU"   "JM"   "NPU"  "GKT"  "OU"
## [61] "CH"   "DZ"   "HJ"   "MP"   "EM"   "HP"   "CP"   "EJ"   "MO"   "B"
## [71] "FP"   "NO"   "NT"   "EV"   "KMT"  "KO"   "HQ"   "MNU"  "CG"   "LO"
## [81] "DT"   "OT"   "KLMT" "MQ"   "GIM"  "DN"   "IL"
```

G: Carbohydrate-transport and metabolism

```
length(which(str_detect(enkel_eggNOG_res$COG_category, pattern = "G")))

## [1] 4023
```

P: Inorganic ion transport and metabolism

```
length(which(str_detect(enkel_eggNOG_res$COG_category, pattern = "P")))

## [1] 2853
```

E: Amino acid transport and metabolism

```
length(which(str_detect(enkel_eggNOG_res$COG_category, pattern = "E")))

## [1] 2511
```

I: Lipid transport and metabolism

```
length(which(str_detect(enkel_eggNOG_res$COG_category, pattern = "I")))

## [1] 699
```

S: Function unknown

```
length(which(str_detect(enkel_eggNOG_res$COG_category, pattern = "S")))

## [1] 6560
```

C5 – K-numbers for the Reconstruction of Pathways

Kegg_ko’s for *Bacteroides* and *Bifidobacterium* were listed separately in excel. The following code was used to remove unwanted subjects, before annotating the K-numbers in KEGG Mapper Reconstruct. They were saved separately as CSV files, which were opened in excel and saved in notebook.

```
KO_bifido <- read_excel("KO_bifido.xlsx")
#which(str_detect(KO_bifido$KEGG_ko, pattern = "ko:"))
KO_bifido$KEGG_ko <- str_remove_all(KO_bifido$KEGG_ko, pattern = "ko:")
KO_bifido$KEGG_ko <- str_replace_all(KO_bifido$KEGG_ko, pattern = ",", repl
acement = ";")
write.csv2(KO_bifido, file = "KO_bifido.csv")

KO_bacteroides <- read_excel("KO_bacteroides.xlsx")
#which(str_detect(KO_bacteroides$KEGG_ko, pattern = "ko:"))
KO_bacteroides$KEGG_ko <- str_remove_all(KO_bacteroides$KEGG_ko, pattern =
"ko:")
KO_bacteroides$KEGG_ko <- str_replace_all(KO_bacteroides$KEGG_ko, pattern =
",", replacement = ";")
write.csv2(KO_bacteroides, file = "KO_bacteroides.csv")
```

Split columns with “;” to separate all K-numbers in excel.

C6 – dbCAN Annotation for Glycoside Hydrolase Families

Started by making an overview of all GH families (171) in a table, from the cazy-website

library(rvest)

```
## Warning: package 'rvest' was built under R version 4.1.2

##
## Attaching package: 'rvest'

## The following object is masked from 'package:readr':
##
##     guess_encoding

library (dplyr)

start_link <- "http://www.cazy.org/GH"
```

```
i <- 1:171
end_link <- ".html"

vector_link <- paste0(start_link, i, end_link, sep = "")
data_frame_genes <- as.data.frame(matrix("NA", ncol = 2, nrow = 171))


for (a in 1:171) {
  link = vector_link[a]

  page = read_html(link)

  name = page %>% html_nodes("tr:nth-child(1) .tdsum") %>% html_text()

  data_frame_genes[a,2] <- name
  data_frame_genes[a,1] <- paste0("GH", a, sep = "")
  #print(a)
}
head(data_frame_genes)

##    V1
## 1 GH1
## 2 GH2
## 3 GH3
## 4 GH4
## 5 GH5
## 6 GH6
#
#
V2
##
1
ß-glucosidase (EC 3.2.1.21); ß-galactosidase (EC 3.2.1.23); ß-mannosidase (
EC 3.2.1.25); ß-glucuronidase (EC 3.2.1.31); ß-xylosidase (EC 3.2.1.37); ß-
D-fucosidase (EC 3.2.1.38); phlorizin hydrolase (EC 3.2.1.62); exo-ß-1,4-gl
ucanase (EC 3.2.1.74); 6-phospho-ß-galactosidase (EC 3.2.1.85); 6-phospho-ß
-glucosidase (EC 3.2.1.86); strictosidine ß-glucosidase (EC 3.2.1.105); lac
tase (EC 3.2.1.108); amygdalin ß-glucosidase (EC 3.2.1.117); prunasin ß-glu
cosidase (EC 3.2.1.118); vicianin hydrolase (EC 3.2.1.119); raucaffricine ß
-glucosidase (EC 3.2.1.125); thioglucosidase (EC 3.2.1.147); ß-primeverosid
ase (EC 3.2.1.149); isoflavonoid 7-O-ß-apiosyl-ß-glucosidase (EC 3.2.1.161)
; ABA-specific ß-glucosidase (EC 3.2.1.175); DIMBOA ß-glucosidase (EC 3.2.1
.182); ß-glycosidase (EC 3.2.1.-); hydroxyisourate hydrolase (EC 3.-.-.-);
ß-rutinosidase /a-L-rhamnose-(1,6)-ß-D-glucosidase (EC 3.2.1.-); protodiosc
in 26-O-Î²-D-glucosidase (EC 3.2.1.186)
##
2
ß-galactosidase (EC 3.2.1.23) ; ß-mannosidase (EC 3.2.1.25); ß-glucuronidas
e (EC 3.2.1.31); a-L-arabinofuranosidase (EC 3.2.1.55); mannosylglycoprotei
n endo-ß-mannosidase (EC 3.2.1.152); exo-ß-glucosaminidase (EC 3.2.1.165);
a-L-arabinopyranosidase (EC 3.2.1.-); ß-galacturonidase (EC 3.2.1.-); ß-xyl
osidase (EC 3.2.1.37); ß-D-galactofuranosidase (EC 3.2.1.146); ß-glucosidas
e (EC 3.2.1.21)
##
```

```
3
ß-glucosidase (EC 3.2.1.21); xylan 1,4-ß-xylosidase (EC 3.2.1.37); ß-glucos
ylceramidase (EC 3.2.1.45); ß-N-acetylhexosaminidase (EC 3.2.1.52); a-L-ara
binofuranosidase (EC 3.2.1.55); glucan 1,4-ß-glucosidase (EC 3.2.1.74); iso
primeverose-producing oligoxyloglucan hydrolase (EC 3.2.1.120); coniferin ß
-glucosidase (EC 3.2.1.126); exo-1,3-1,4-glucanase (EC 3.2.1.-); ß-N-acetyl
glucosaminide phosphorylases (EC 2.4.1.-); ß-1,2-glucosidase (EC 3.2.1.-);
ß-1,3-glucosidase (EC 3.2.1.-); xyloglucan-specific exo-ß-1,4-glucanase / e
xo-xyloglucanase (EC 3.2.1.155); stevioside-ß-1,2-glucosidase (EC 3.2.1.-);
lichenase / endo-ß-1,3-1,4-glucanase (EC 3.2.1.73); protodioscin 26-O-ß-D-g
lucosidase (EC 3.2.1.186); ß-glucuronidase (EC 3.2.1.31)
##
4
maltose-6-phosphate glucosidase (EC 3.2.1.122); a-glucosidase (EC 3.2.1.20)
; a-galactosidase (EC 3.2.1.22); 6-phospho-ß-glucosidase (EC 3.2.1.86); a-g
lucuronidase (EC 3.2.1.139); a-galacturonase (EC 3.2.1.67); palatinase (EC
3.2.1.-)
## 5 endo-ß-1,4-glucanase / cellulase (EC 3.2.1.4); endo-ß-1,4-xylanase (EC
3.2.1.8); ß-glucosidase (EC 3.2.1.21); ß-mannosidase (EC 3.2.1.25); ß-gluco
sylceramidase (EC 3.2.1.45); glucan ß-1,3-glucosidase (EC 3.2.1.58); exo-ß-
1,4-glucanase / cellodextrinase (EC 3.2.1.74); glucan endo-1,6-ß-glucosidas
e (EC 3.2.1.75); mannan endo-ß-1,4-mannosidase (EC 3.2.1.78); cellulose ß-1
,4-cellobiosidase (EC 3.2.1.91); steryl ß-glucosidase (EC 3.2.1.104); endog
lycoceramidase (EC 3.2.1.123); ß-primeverosidase (EC 3.2.1.149); xyloglucan
-specific endo-ß-1,4-glucanase (EC 3.2.1.151); endo-ß-1,6-galactanase (EC 3
.2.1.164); ß-1,3-mannanase (EC 3.2.1.-); arabinoxylan-specific endo-ß-1,4-x
ylanase (EC 3.2.1.-); mannan transglycosylase (EC 2.4.1.-); lichenase / end
o-ß-1,3-1,4-glucanase (EC 3.2.1.73); ß-glycosidase (EC 3.2.1.-); endo-ß-1,3
-glucanase / laminarinase (EC 3.2.1.39); ß-N-acetylhexosaminidase (EC 3.2.1
.52); chitosanase (EC 3.2.1.132); ß-D-galactofuranosidase (EC 3.2.1.146); ß
-galactosylceramidase (EC 3.2.1.46); ; ß-rutinosidase /a-L-rhamnose-(1,6)-ß
-D-glucosidase (EC 3.2.1.-); a-L-arabinofuranosidase (EC 3.2.1.55); glucoma
nnan-specific endo-ß-1,4-glucanase (EC 3.2.1.-); hesperidin 6-O-a-L-rhamnos
yl-ß-glucosidase (EC 3.2.1.168)
##
6
endoglucanase (EC 3.2.1.4); cellobiohydrolase (EC 3.2.1.91); lichenase / en
do-ß-1,3-1,4-glucanase (EC 3.2.1.73);
```

dbCAN had a limit with files no larger than 20MB. The eggNOG-file was almost 40MB and had to be split into two separate files for dbCAN-annotation and then put back together again.

```
eggNOG_fil_1 <- eggNOG_fil[1:76548]
eggNOG_fil_2 <- eggNOG_fil[76549:153094]
write.csv2(eggNOG_fil_1, file = "dbCAN_del1.csv") #Aminoacid sequence 1 and
2
write.csv2(eggNOG_fil_2, file = "dbCAN_fil2.csv")

library(readxl)
dbCANresultDel1 <- read_xlsx("dbCANresultDel1.xlsx")
dbCANresultDel2 <- read_xlsx("dbCANresultdel2.xlsx")

dbCANresults <- rbind(dbCANresultDel1, dbCANresultDel2) #Bound together
```

All data from the dbCAN metaserver was uploaded in excel, and all proteins with only one tool-match were removed for all nodes. The excel-sheets were then uploaded to Rstudio, and the correct GHs were put in the correct cell.

Wanted to remove unnecessary information and signs from the table.

```r
dbCANresults$HMMER <- gsub("\\s*\\([^\\)]+\\)", "", dbCANresults$HMMER)

#Fjerner rekken jeg ikke trenger med "signalP"
dbCANresults <- dbCANresults[,-6]

#Only HMMER-proteins, without the "_"
listeBact <- strsplit(dbCANresults$HMMER, split = "_")

#Replacing all HMMER-values with HMMER-values without the "_" using a for-l
oop
for (i in 1:nrow(dbCANresults)) {
  Variabel_bact <- listeBact[[i]][1]
  dbCANresults$HMMER[i] <- Variabel_bact
}

#And did the same with DIAMOND-values
dbCANresults$DIAMOND <- gsub("\\s*\\([^\\)]+\\)", "", dbCANresults$DIAMOND)

listeBact2 <- strsplit(dbCANresults$DIAMOND, split = "_")
for (i in 1:nrow(dbCANresults)) {
  variabel_bact2 <- listeBact2[[i]][1]
  dbCANresults$DIAMOND[i] <- variabel_bact2
}
```

Made a new column meant for those GHs that were equal for at least two tools used in dbCAN.

```r
dbCANresults$equalAtleast2 <- "NA"
```

GHs matching will be put into the new column with this code: HMMER vs eCAMI

```r
indeks_HMMER_eCAMI <- which(dbCANresults$HMMER == dbCANresults$eCAMI)
dbCANresults$equalAtleast2[indeks_HMMER_eCAMI] <- dbCANresults$HMMER[indeks
_HMMER_eCAMI]

#Doublechecking:
indeks_NA <- which(dbCANresults$equalAtleast2 == "NA")
length(indeks_NA)

## [1] 1776
```

HMMER vs DIAMOND Using the indeks_Na to fill in

```r
indeks_HMMER_DMND <- which(dbCANresults$HMMER[indeks_NA] == dbCANresults$DI
AMOND[indeks_NA])
limes_inn <- indeks_NA[indeks_HMMER_DMND]
```

```
dbCANresults$equalAtleast2[limes_inn] <- dbCANresults$DIAMOND[limes_inn]
#Doublechecking:
indeks_NA2 <- which(dbCANresults$equalAtleast2 == "NA")
length(indeks_NA2)

## [1] 860
```

eCAMI vs DIAMOND Using the Indeks_NA2 to fill in

```
indeks_eCAMI_DMND <- which(dbCANresults$eCAMI[indeks_NA2] == dbCANresults$D
IAMOND[indeks_NA2])
limes_inn_2 <- indeks_NA2[indeks_eCAMI_DMND]
dbCANresults$equalAtleast2[limes_inn_2] <- dbCANresults$eCAMI[limes_inn_2]
#Doublechecking:
indeks_NA3 <- which(dbCANresults$equalAtleast2 == "NA")
length(indeks_NA3)

## [1] 230
```

Cells that still have NA-values could be a result of “+” values or different orders on several GHs.

```
Kun_ett_verktoy <- which(dbCANresults$`#ofTools` == "1")
dbCANresults <- dbCANresults[-Kun_ett_verktoy,]
```

Wanted to remove + values

```
test123 <- which(dbCANresults$equalAtleast2 == "NA")

#Checking that none are the same
which(dbCANresults$HMMER[test123] == dbCANresults$eCAMI[test123])

## integer(0)

#Lager en vektor med de som har NA-verdier i HMMER. Fjerner pluss-tegn og s
plitter
vektor_HMMER <- dbCANresults$HMMER[test123]
HMMER_split <- strsplit(vektor_HMMER, split = "\\+")
HMMER_split[[5]][1] #Sjekker

## [1] "GT4"

#For DIAMOND
Vektor_DIAMOND <- dbCANresults$DIAMOND[test123]
DIAMOND_split <- strsplit(Vektor_DIAMOND, split = "\\+")
DIAMOND_split[[3]][1] #Sjekker

## [1] "CBM35"

#For eCAMI
Vektor_eCAMI <- dbCANresults$eCAMI[test123]
eCAMI_split <- strsplit(Vektor_eCAMI, split = "\\+")
eCAMI_split[[228]][3]

## [1] "CBM13"
```

Got the GHs that matched between eCAMI and DIAMOND

```
library(stringr)
for (i in 1:length(test123)) {
  b <- 1:length(eCAMI_split[[i]])
  c <- 1:length(DIAMOND_split[[i]])
  c1 <- ifelse(length(which(eCAMI_split[[i]][b] == DIAMOND_split[[i]][c[1]]
)) > 0, 1, 0)
  c2 <- ifelse(length(which(eCAMI_split[[i]][b] == DIAMOND_split[[i]][c[2]]
)) > 0, 1, 0)
  c3 <- ifelse(length(which(eCAMI_split[[i]][b] == DIAMOND_split[[i]][c[3]]
)) > 0, 1, 0)
  c4 <- ifelse(length(which(eCAMI_split[[i]][b] == DIAMOND_split[[i]][c[4]]
)) > 0, 1, 0)
  c5 <- ifelse(length(which(eCAMI_split[[i]][b] == DIAMOND_split[[i]][c[5]]
)) > 0, 1, 0)

  c_all <- c(c1,c2,c3,c4,c5)
  c_all_2 <- which(c_all > 0, arr.ind = FALSE)
  c_all_3 <- paste(DIAMOND_split[[i]][c_all_2], sep = ",")
  c_all_4 <- str_c(c_all_3, collapse = "+")
  dbCANresults$equalAtleast2[test123[i]] <- c_all_4
}
which(dbCANresults$equalAtleast2 == "NA") # Ingen verdier med NA
```

```
## integer(0)
```

```
head(which(dbCANresults$equalAtleast2 == ""))
```

```
## [1]   3  11  15  18 101 123
```

Did the exact same thing with HMMER and DIAMOND

```
test1234 <- which(dbCANresults$equalAtleast2 == "")
which(dbCANresults$HMMER[test1234] == dbCANresults$DIAMOND[test1234])
```

```
## integer(0)
```

```
library(stringr)
for (i in 1:length(test1234)) {
  e <- 1:length(HMMER_split[[i]])
  d <- 1:length(DIAMOND_split[[i]])
  d1 <- ifelse(length(which(HMMER_split[[i]][e] == DIAMOND_split[[i]][d[1]]
)) > 0, 1, 0)
  d2 <- ifelse(length(which(HMMER_split[[i]][e] == DIAMOND_split[[i]][d[2]]
)) > 0, 1, 0)
  d3 <- ifelse(length(which(HMMER_split[[i]][e] == DIAMOND_split[[i]][d[3]]
)) > 0, 1, 0)
  d4 <- ifelse(length(which(HMMER_split[[i]][e] == DIAMOND_split[[i]][d[4]]
)) > 0, 1, 0)
  d5 <- ifelse(length(which(HMMER_split[[i]][e] == DIAMOND_split[[i]][d[5]]
)) > 0, 1, 0)

  d_all <- c(d1,d2,d3,d4,d5)
  d_all_2 <- which(d_all > 0, arr.ind = FALSE)
  d_all_3 <- paste(DIAMOND_split[[i]][d_all_2], sep = ",")
  d_all_4 <- str_c(d_all_3, collapse = "+")
  dbCANresults$equalAtleast2[test1234[i]] <- d_all_4
}
```

And with eCAMI and HMMER

```
test12345 <- which(dbCANresults$equalAtleast2 == "")
which(dbCANresults$HMMER[test12345] == dbCANresults$DIAMOND[test12345])

## integer(0)

library(stringr)
for (i in 1:length(test12345)) {
  e <- 1:length(HMMER_split[[i]])
  d <- 1:length(eCAMI_split[[i]])
  d1 <- ifelse(length(which(HMMER_split[[i]][e] == eCAMI_split[[i]][d[1]]))
> 0, 1, 0)
  d2 <- ifelse(length(which(HMMER_split[[i]][e] == eCAMI_split[[i]][d[2]]))
> 0, 1, 0)
  d3 <- ifelse(length(which(HMMER_split[[i]][e] == eCAMI_split[[i]][d[3]]))
> 0, 1, 0)
  d4 <- ifelse(length(which(HMMER_split[[i]][e] == eCAMI_split[[i]][d[4]]))
> 0, 1, 0)
  d5 <- ifelse(length(which(HMMER_split[[i]][e] == eCAMI_split[[i]][d[5]]))
> 0, 1, 0)

  d_all <- c(d1,d2,d3,d4,d5)
  d_all_2 <- which(d_all > 0, arr.ind = FALSE)
  d_all_3 <- paste(HMMER_split[[i]][d_all_2], sep = ",")
  d_all_4 <- str_c(d_all_3, collapse = "+")
  dbCANresults$equalAtleast2[test12345[i]] <- d_all_4
}
```

A few cells did not have any values at all. These were checked up manually and fixed.

```
which(dbCANresults$equalAtleast2 == "N")

## integer(0)

which(dbCANresults$equalAtleast2 == "")

##  [1]   11  555  607  660  929 1244 1580 1870 2010 2036 2039 2212 2668 26
81 2711
## [16] 2764 2792 2865 2898 2950 3265 3443 3554 3601 3782 3895 3915

dbCANresults$equalAtleast2[102] <- "GH20"
dbCANresults$equalAtleast2[329] <- "GH171"
dbCANresults$equalAtleast2[719] <- "GH2"
dbCANresults$equalAtleast2[1278] <- "GH2"
dbCANresults$equalAtleast2[1455] <- "GH43"
dbCANresults$equalAtleast2[1824] <- "GH30"
dbCANresults$equalAtleast2[1825] <- "GH154"
dbCANresults$equalAtleast2[1837] <- "GH2"
dbCANresults$equalAtleast2[2036] <- "GH13"
```

```
dbCANresults$equalAtleast2[2972] <- "GH43"
dbCANresults$equalAtleast2[3111] <- "GH78"
dbCANresults$equalAtleast2[3128] <- "GH2"
```

The rest of the blanks did not have any matches and was removed.

```
slette <- which(dbCANresults$equalAtleast2 == "")
library(stringr)
library(dplyr)
dbCANresults <- dbCANresults[-slette,]

length(unique(dbCANresults$equalAtleast2))

## [1] 197
```

Finished the table and checked the GHs present

```
dbCANfinal_result <- dbCANresults[,-c(3:6)]
write.csv2(dbCANfinal_result, file = "dbCANfinal_result.csv")
```

dbCANfinal_result <- read.csv2(“dbCANfinal_result.csv”)

Put all GH-annotations in the already filtered (by e-values) EggNOG-table, to have it all in one place.

```
EggNOG_gene_taxonomy <- read.csv2("EggNOG_gene_taxonomy.csv")
```

Made a new column in the table

```
GH_nodes <- read_excel("GH_nodes.xlsx")

EggNOG_gene_taxonomy$GH_dbCAN <- "NA"
```

Put the GHs from dbCAN-table into eggNOG-table.

```
           for (i in 1:nrow(EggNOG_gene_taxonomy)) {
             eggnog_match <- which(EggNOG_gene_taxonomy$accn == GH_nodes$`G
ene ID`[i])
             GH_family_name <- GH_nodes$equalAtleast2.1[i]
             EggNOG_gene_taxonomy$GH_dbCAN[eggnog_match] <- GH_family_name
           }
```

Made a separate table with only GHs that were filtered by e-values.

```
library(stringr)
alle_GH_genes <- which(str_detect(EggNOG_gene_taxonomy$Kolonne2, pattern =
"GH"))
kun_gh_gener <- EggNOG_gene_taxonomy[alle_GH_genes,]
kun_gh_gener <- as.matrix(kun_gh_gener)

write.csv2(kun_gh_gener, file = "GH_genes.csv")
```

To load the file again: Kun_gh_gener <- read.csv2(“Alle_gh_gener_eggNOG_cazy.csv”)

C7 – Statistical Analysis of Short-Chain Fatty Acids

Correlation plot of SCFAs and 16S taxonomy

In excel, a matrix was made with the relative values of the abundance (0-1) of bacterial taxonomies from the 16S data, and the relative values of short-chain fatty acid concentration presented in mmol/kg feces.

```
library(readxl)
My.data <- read_excel("Corrplot.xlsx")

## New names:
## * Other -> Other...33
## * Other -> Other...37
```

Creating a correlation matrix using spearman correlation, showing correlation between variables.

```
My.corr.data <- cor(My.data, method = c("spearman"))
```

Looking at significant p-values, which will be used to make the correlation plot.

```
library(Hmisc)

My.rcorr.data <- rcorr(as.matrix(My.data))

mydata.coeff <- My.rcorr.data$r
mydata.p <- My.rcorr.data$P

#signif.mydata <- p.adjust(mydata.p, method = "hochberg")
#signif.mydata <- as.data.frame(signif.mydata)
```

Visualizing the plot, with the p significance level 0.1.

```
corrplot(My.corr.data, method = "circle", type = "lower", p.mat = mydata.p,
sig.level = 0.1, insig = "blank", tl.col = "black")
```

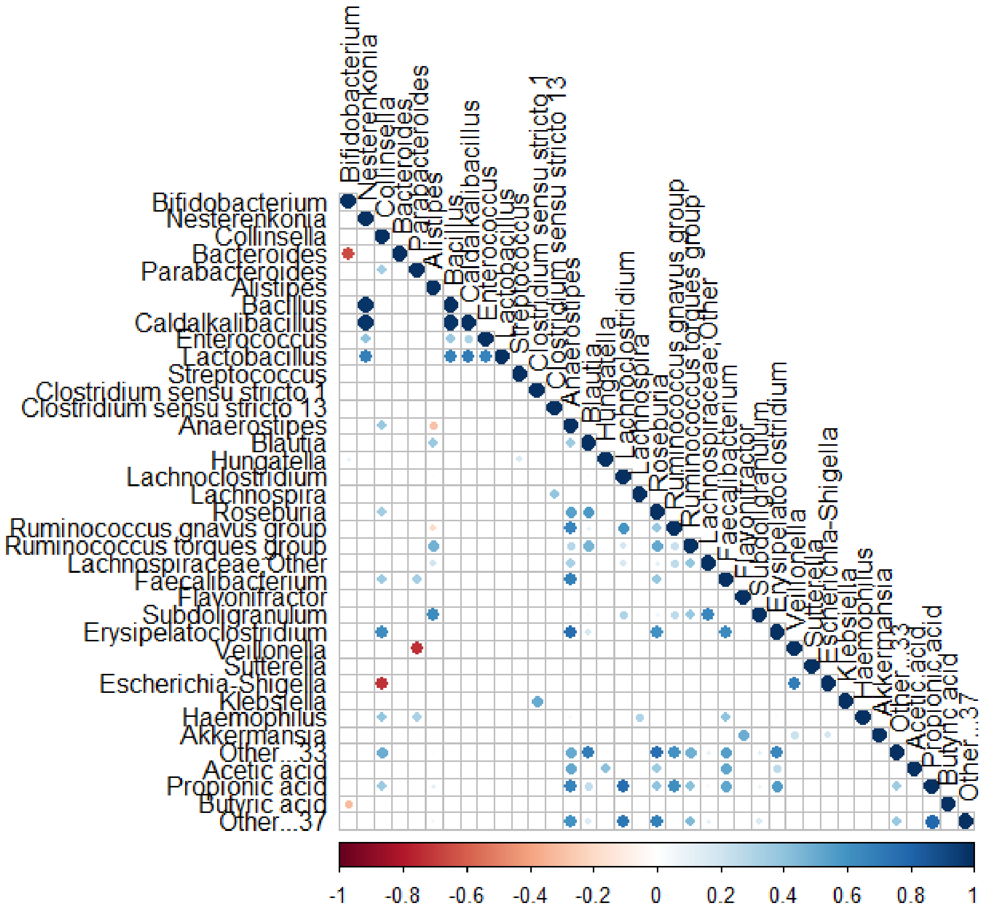

Figure C7.1. Correlation plot between genus and SCFA concentration. P-value < 0.1 with spearman correlation. Red circles indicate negative correlation, blue circles indicate positive correlation. The darker the colour and the bigger the circle, the bigger correlation is there. The figure is made in R-studio in cooperation with Marianne Frøseth.

Two additional correlation plots were made but investigating the groups separately (high in *Bacteroides* and high in *Bifidobacterium*). The same method as illustrated above was used, only with a p-value < 0.05. Still, no correlations between *Bacteroides* or *Bifidobacterium* and SCFAs appeared.

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

Postboks 5003
NO-1432 Ås
Norway