Norwegian University
of Life Sciences

Master's Thesis 2021   60 ECTS
Faculty of Chemistry, Biotechnology and Food Science

The potential for Human Milk Oligosaccharide utilization by *Bifidobacterium* in 6 months children

Tonje Nilsen
MSc Biotechnology

The potential for Human Milk Oligosaccharide utilization by *Bifidobacterium* in 6 months children

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 Norwegian University of Life Sciences, under the supervision of Professor Knut Rudi and Ph.D. Morten Nilsen.

Firstly, I would like to thank my main supervisor, Knut Rudi, and my secondary supervisor, Morten Nilsen. Your enthusiasm, clever ideas, solutions to problems and all help throughout the writing process has been greatly appreciated. A special thanks to Morten Nilsen for all help you have provided in relation to laboratory work, bioinformatic analyzes and data interpretation, and for your never-ending positive attitude. I would also like to thank the whole MiDiv group for making me feel included and for all the knowledge you have shared throughout the year.

I would like to thank the PreventADALL group for giving me the opportunity to analyze their samples, and have the information connected to them.

Furthermore, I would like to thank the other master students in the group. I appreciate all the social evenings, lunches and the good company in the laboratory, and this year would not have been the same without you. I would also like to thank my wonderful roommates for always being positive and giving me motivation during this year. I could not have asked for better roommates.

Finally, a big thanks friends and family for all the love, support and encouragement you have provided, not only through this year, but throughout my studies.

Ås, June 2021
Tonje Nilsen

Abstract

The infant gut microbiota is known to be dominated by *Bifidobacterium*, especially in healthy, breastfed infants. This is mainly due to their ability to utilize human milk oligosaccharides (HMOs) that are non-digestible glycans, unique to breast milk. From this utilization, metabolites such as short chain fatty acids (SCFAs) are produced, that have an important role in nurturing the epithelial cells in the large intestine. There is currently a knowledge gap related to how *Bifidobacterium* utilize HMOs in the infant gut. The aim of this thesis was therefore to analyze how *Bifidobacterium* degrade HMOs in the infant gut using a metagenomic and proteomic approach.

Potential HMO degradation by *Bifidobacterium* was studied using fecal samples from the PreventADALL study. To obtain an overview of the gut microbiota composition, and to select samples with high abundance of *Bifidobacterium* for further analyzes, a 16S rRNA sequencing was performed. The detailed composition and functional potential of *Bifidobacterium* species was found through a shotgun sequencing. To identify HMO utilizing proteins found in *Bifidobacterium*, a proteome analysis was performed, and the proteins were divided into different HMO degradation pathways. Several proteins related to HMO degradation were found either from both the shotgun and proteome data, or only from the shotgun data. For three out of five building blocks of HMO, whole degradation pathways were found. In addition to this, all the main enzymes to break down HMO; β-galactosidase, fucosidase, sialidase, GLNBP and β-hexosaminidase, were identified from the data.

In conclusion, *Bifidobacterium* has the ability to degrade HMO compounds, and there is a high potential that some *Bifidobacterium* species contain whole HMO degradation pathways. This provides a good base to research different HMO degradation pathways in specific *Bifidobacterium* species.

Sammendrag

Tarmmikrobiotaen til spedbarn er kjent for å være dominert av *Bifidobacterium*, spesielt hos friske, ammede spedbarn. Dette er hovedsakelig grunnet deres egenskaper til å utnytte spesifikke oligosakkarider (HMOer) i morsmelk, som er ikke-nedbrytbare glykaner. Fra denne nedbrytelsen blir det produsert metabolitter, slik som kortkjedede fettsyrer (SCFAer), som har en viktig rolle i å fungere som næring for epitelceller i tykktarmen. Det er for øyeblikket mangel på kunnskap relatert til hvordan *Bifidobacterium* bryter ned HMOer i tarmen til spedbarn. Målet med denne oppgaven var derfor å analysere hvordan *Bifidobacterium* bryter ned HMOer i tarmen til spedbarn ved å bruke en metagenomisk og proteomisk analyse.

Potensiell HMO-nedbrytelse av *Bifidobacterium* ble studert ved å bruke avføringsprøver hentet fra PreventADALL-studien. For å få en oversikt over tarmmikrobiota-sammensetningen, og for å velge ut prøver med høy tilstedeværelse av *Bifidobacterium* for videre analyser, ble det utført en 16S rRNA sekvensering. Den detaljerte sammensetningen og det funksjonelle potensialet av *Bifidobacterium*-arter ble funnet gjennom en shotgun-sekvensering. For å identifisere HMO-nedbrytende proteiner funnet i *Bifidobacterium* ble det utført en proteom-analyse, og proteinene ble delt inn i ulike HMO-nedbrytende veier. Flere proteiner relatert til HMO-nedbrytelse ble funnet enten fra både shotgun- og proteom-dataene, eller bare fra shotgun-dataene. For tre av fire byggeklosser i HMO ble det funnet fullstendige nedbrytelsesveier. I tillegg til dette ble alle hovedenzymene som bryter ned HMO: β-galaktosidase, fukosidase, sialidase, GLNBP og β-heksosaminidase, identifisert fra dataene.

For å konkludere har *Bifidobacterium* egenskapen til å bryte ned komponenter av HMO, og det er et høyt potensial for at noen *Bifidobacterium*-arter inneholder fullstendige HMO-nedbrytende veier. Dette gir et godt grunnlag for å undersøke ulike HMO-nedbrytende veier i spesifikke *Bifidobacterium*-arter.

Abbreviations

ABC transporter	ATP-binding cassette transporter
ATP	Adenosine triphosphate
cDNA	complementary DNA
DNA	Deoxyribonucleic acid
F6PPK	Fructose-6-phosphate phosphoketolase
Fuc	Fucose
Gal	Galactose
GalE	UDP-glucose/GlcNAc 4-epimerase
GalK	Galactokinase
GalNAc	N-acetylgalactosamine
GalT	UDP-glucose-hexose-1-phosphate uridylyl transferase
GC	Gas chromatography
Glc	Glucose
GlcNAc	N-acetylglucosamine
GNB	Galacto-N-biose
GLNBP	GNB/LNB phosphorylase
HMO	Human milk oligosaccharide
Lac	Lactose
LC-MS/MS	Liquid chromatography-tandem mass spectrometry
LNB	Lacto-N-biose
LNT	Lacto-N-tetraose
mRNA	messenger RNA
NahK	N-acetylhexosamine-1-kinase
Neu5Ac	N-acetyl neuraminic acid
P	Phosphate
PCR	Polymerase chain reaction
qPCR	Quantitative PCR
RNA	Ribonucleic acid
rRNA	ribosomal RNA
SCFA	Short chain fatty acid
SDS-PAGE	Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis

Table of Contents

1 Introduction

1.1 Infant and adult-like gut microbiota and colonization

Humans have almost equal amounts of cells and bacteria in the body, with the highest density of bacteria in the large intestine (Thursby & Juge, 2017). The complex community of microorganisms in the intestine is referred to as the gut microbiota and has evolved to survive in the gastrointestinal tract (Milani et al., 2017). The gut microbiota can consist of harmless commensals, opportunistic pathogens or health promoting microorganisms (O'Callaghan & van Sinderen, 2016). In the large intestine it is discovered approximately 160 species that creates the gut microbiota (Rodriguez et al., 2015). The bacteria composition varies according to environmental factors in the gut, such as pH, temperature, access to oxygen, nutrients available and so on. The bacteria that survive the gut environment best will often dominate (Milani et al., 2017).

The composition of bacteria in the gut can give indications of different conditions, for example dysbiosis (Olin et al., 2018), which is an continuous imbalance in the gut microbiota, due to changes in the composition and metabolic activity (Belizário & Faintuch, 2018). Dysbiosis can lead to long term effects like obesity, diabetes and inflammatory bowel disease (IBD) (Milani et al., 2017). Diseases like IBD and psoriatic arthritis are both linked to a loss of diversity in the gut microbiota (Morrison & Preston, 2016), and this is just a few examples of several studies that have linked development of different diseases due to an altered gut microbiota (Arboleya et al., 2016).

Humans are dependent on the gut microbiota, because the bacteria break down different food compounds that we cannot digest, which results in the production of short chain fatty acids (SCFAs), that nurture the epithelial cells in the large intestine. The gut microbiota also protects the intestinal cells from pathogen colonization and helps mature the immune system (Milani et al., 2017).

The infant gut microbiota starts from birth and consists mainly of some bacteria from the mother and environment. From there the diversity in the microbiota increases, and a child have an adult-like microbiota at the age of 3-5 years old (Rodriguez et al., 2015). From the stage of newborn to a child of age 3-5 there are a large number of factors that can affect the microbiota diversity and composition. During pregnancy, factors like maternal microbiota, health status and lifestyle can affect the child. During birth will the mode of delivery, vaginal vs c-section, term vs preterm and antibiotic treatment have an impact on the infant gut microbiota. Breastfeeding vs formula, genetics, duration of lactation, family environment and geographical location will all modulate the infant gut microbiota during the first few years of life (O'Callaghan & van Sinderen, 2016; Rodriguez et al., 2015). During the adult-period of life, mainly lifestyle and diet are the factors that can modulate the gut microbiota, and as an elder, living conditions and medications are important factors (Rodriguez et al., 2015).

The meconium, which is the first stool of an infant, is not sterile and consists of a community of microorganisms with Firmicutes as the main phylum and staphylococci as the dominant bacteria (Rodriguez et al., 2015). The fist colonizers of the infant gut create an environment that fit strict anaerobic bacteria, such as *Bacteroides*, *Clostridium* and *Bifidobacterium*. The newborn gut microbiota is known to have low diversity and is dominated by the phyla Proteobacteria and Actinobacteria. New phyla dominate during increasing time after birth, and these phyla are Firmicutes and Bacteroidetes (Rodriguez et al., 2015). *Bifidobacterium* will dominate the gut microbiota in healthy, breastfed children (O'Callaghan & van Sinderen, 2016), and contribute to more than 50% of the total bacteria population during the breast feeding period (Gotoh et al., 2018). This domination is mainly due to their ability to utilize human milk oligosaccharides (HMOs) found in human breast milk. During weaning the population of *Bifidobacterium* decreases.

1.2 Human milk oligosaccharides

Human milk oligosaccharides (HMOs) are a family with structurally different glycans, that are unique to breast milk (Bode, 2012). They are non-digestible oligosaccharides that are metabolized by gut bacteria in the large intestine (O'Callaghan & van Sinderen, 2016), and therefore have a major part in shaping the gut microbiota in breastfed infants (Bode, 2012). One bacterium in particular that is common in breastfed infants is *Bifidobacterium* (Kitaoka, 2012). It is discovered over 130 different oligosaccharides in breast milk that are HMOs, which makes it a complex composition (Bode, 2006; Kitaoka, 2012). According to (Bode, 2012), HMOs are antimicrobial agents that can prevent adhesion. They also work as soluble decoy receptors, which means they can recognize and bind to specific growth factors or cytokines but cannot send signals or activate receptor complexes (Mantovani et al., 2001). The human milk oligosaccharides also prevent pathogens to adhere to the infant’s mucosal surfaces, and therefore reduce the risk of infections. It is suggested that HMOs have a prebiotic effect (Bode, 2012). Prebiotic agents are resistant to gastric acid, utilization from the hosts enzymes and gastrointestinal absorption (Davani-Davari et al., 2019). HMOs cover all these factors. The breast milk is not sterile, and it is also seen that HMOs can have positive effects on the lactating mother. HMOs are for example discovered in the mothers urine right before birth, and this can indicate potential positive systemic effects on the mother (Bode, 2012).

The building blocks of HMOs are the monosaccharides D-glucose (Glc), D-galactose (Gal), N-acetylglucosamine (GlcNAc), L-fucose (Fuc) and sialic acid (Sia), the latter is often in the form as N-acetyl neuraminic acid (Neu5Ac) (Bode, 2012). All HMOs have lactose (Gal(β1-4)Glc) at the reducing end, and this can be elongated by disaccharides in either a type 1 or type 2 chain (Figure 1.1). A type 1 chain consists of β1-3 or β1-6 bound lacto-N-biose (LNB, Gal(β1-3)GlcNAc), and this will terminate the chain, which means the core HMO structure is lacto-N-tetraose (LNT, Gal(β1-3)GlcNAc(β1-3)Gal(β1-4)Glc) (Sela et al., 2008). A type 2 chain consists of a β1-3 or β1-6 bound N-acetyllactosamine (Gal(β1-4)GlcNAc), and this can further be elongated by one of the two disaccharides. Lactose or the elongated chain can be fucosylated or sialylated with different α-bonds (Bode, 2012). Some examples of neutral HMOs, which are neither fucosylated or sialylated, are lacto-N-tetraose (LNT), lacto-N-neotetraose (LNnT) and lacto-N-hexaose (LNH). 2-fucosyllactose (2FL) and 3-fucosyllactose (3FL) are examples of fucosylated HMOs, and 3-sialyllactose (3SL) and 6-sialyllactose (6SL) are examples of sialylated HMOs (Garrido et al., 2015).

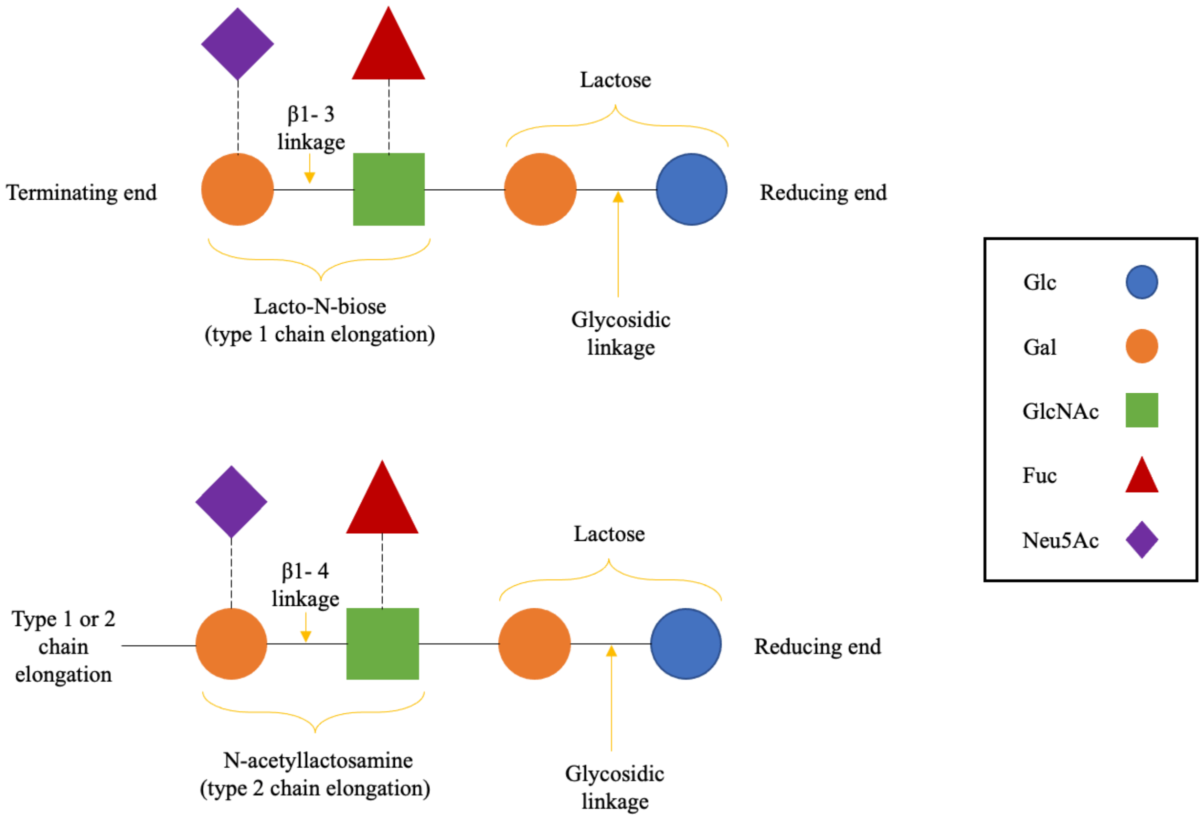


Figure 1.1: Structure of human milk oligosaccharides. The figure shows the structure of HMOs and their building blocks. The dotted lines represent the potential for fucosylation or sialylation. The upper structure shows elongation with type 1 chain, and the lower structure shows elongation with type 2 chain. Glc, glucose; Gal, galactose; GlcNAc, N-acetylglucosamine; Fuc, fucose; Neu5Ac, N-acetyl neuraminic acid (sialic acid). The figure is made based on information from (Bode, 2012).

1.3 *Bifidobacterium*

Bifidobacterium is a gram-positive bacteria genus, belonging to the *Bifidobacteriaceae* family. This family belongs to the phylum Actinobacteria which is known to include bacteria with high GC DNA content. *Bifidobacterium* was originally named *Bacillus bifidus* and classified in the genus *Lactobacillus*, when they were first discovered from feces of a breastfed infant in the late 1800s (Turroni et al., 2011). *Bifidobacterium* is most often found in the gastrointestinal system and is known to be dominating the intestine of healthy, breastfed infants. When the infant ages, the levels reduces, and the composition of *Bifidobacterium* species changes (Arboleya et al., 2016). The most common *Bifidobacterium* species found in the infant gut are *Bifidobacterium longum* subsp. *infantis* and *Bifidobacterium bifidum* (Bunesova et al., 2016), but *Bifidobacterium longum* subsp. *longum* and *Bifidobacterium breve* are also present at a high level (Arboleya et al., 2016). An adult gut microbiota consists more of the *Bifidobacterium* species *Bifidobacterium catenulatum*, *Bifidobacterium adolescentis* and *B. longum* subsp. *longum* (Arboleya et al., 2016).

Bifidobacterium is thought to be vertically transferred from mother to child, by transmission from the vaginal tract during vaginal birth, the gastrointestinal tract and breast milk (Collado et al., 2016; Makino et al., 2013). Therefore, birth mode (vaginally vs. c-section) and to some extent breast feeding will have an impact on *Bifidobacterium* colonization in the infant gut (Dominguez-Bello et al., 2010; Guaraldi & Salvatori, 2012).

The bacteria has GRAS status (generally recognized as safe) and researched health benefits, that makes it a probiotic microorganism (O'Callaghan & van Sinderen, 2016). They are most likely able to produce short chain fatty acids (SCFAs) and bacteriocins, which are health promoting metabolites (Arboleya et al., 2016). *Bifidobacterium* is also important in stimulating the immune system (Arboleya et al., 2016), and as other beneficial gut bacteria, they occupy place and food resources that prevent the growth of pathogenic bacteria in the intestine (Kitaoka, 2012). According to (Underwood et al., 2015), *B. longum* subsp. *infantis* is associated with the ability to decrease intestinal permeability and has anti-inflammatory properties. Since some diseases are linked with altered gut microbiota, several studies have focused on changed levels or composition of *Bifidobacterium* in connection with diseases. Studies have suggested that patients with diseases such as obesity and long-term asthma, also have reduced levels of *Bifidobacterium* (Gao et al., 2015; Hevia et al., 2016). An article by (Di Gioia et al., 2014) has summarized various studies researching the effect of *Bifidobacterium* on diseases such as allergies, celiac disease, obesity, diarrheas, colic and necrotizing enterocolitis.

Bifidobacterium is very common especially in breastfed infants, due to their ability to utilize different components in breast milk. In terms of a *Bifidobacterium* growth factor in breast milk, human milk oligosaccharides (HMOs) are the most promising candidate (Kitaoka, 2012).

1.4 HMO utilization in *Bifidobacterium* species

A whole genome sequencing done on *B. longum* subsp. *infantis* presented gene clusters controlling the expression of glycosidases, sugar transporters and glycan binding proteins specific to HMO utilization (Sela et al., 2008). *B. longum* subsp. *infantis* is also able to grow with HMOs as the only carbon source. *B. bifidum* grow somewhat slower with HMOs as only carbon source and are not able to decompose all monosaccharides from HMOs. In contrast, *B. longum* and *B. breve* alone are hard to grow with HMOs as the only carbon source. This is due to their ability to only utilize some HMOs, but they can catabolize carbohydrates already decomposed by other bacteria (Sela et al., 2008).

Bifidobacterium take up carbohydrates through three different mechanisms; ABC transporters, major facilitator superfamily permeases and phosphotransferase systems (PTS), although the first mechanism is mostly used (Sela et al., 2008). ABC transporters can transport HMO, lactose (Lac), LNB, N-acetylglucosamine (GlcNAc) and sialic acid (often Neu5Ac). The permeases can transport fucose (Fuc), glucose (Glc), galactose (Gal) and Lac, and PTS transport Glc and GlcNAc (figure 1.2).

A 43 kbp gene cluster (Blon_2331 – Blon_2361) has been discovered in *Bifidobacterium* species, mainly *B. longum* subsp. *infantis*. This gene cluster is associated with HMO import and processing (Sela et al., 2008). Some enzymes in this gene cluster are; 1,2-α-fucosidase, 1,3/4-α-fucosidase, 2,3/6-α-sialidase, β-galactosidase and β-N-acetylhexosaminidase (Kitaoka, 2012), and their function is shown in figure 1.2. According to (Matsuki et al., 2016), *Bifidobacterium* species has developed two different ways to break down HMOs. The first way uses extracellular glycoside hydrolases (GH) to break down HMOs to mono- and disaccharides, before incorporating into the cell. The second way is depending on oligosaccharide transporters that import intact HMOs which will be hydrolyzed by intracellular enzymes. *B. bifidum* and some *B. longum* are thought belonging to the group using extracellular hydrolases, whereas some *B. longum*, *B. breve* and *B. longum* subsp. *infantis* belongs to the group using intracellular hydrolases (Odamaki et al., 2015). The extracellular hydrolysis done by *B. bifidum* makes it possible for other (bifido)bacteria to utilize HMO-derivates. This sharing of nutrients is an activity called cross-feeding (Turroni et al., 2018).

Most *Bifidobacterium* species that are common in infants, such as *B. bifidum* and *B. longum* subsp. *infantis*, uses specific enzymes to metabolize galacto-N-biose (GNB) and lacto-N-biose (LNB) (Kitaoka, 2012). LNB is found at the terminating end in HMO structures (Bode, 2012), and is therefore necessary to break down in order to break down HMO. GNB is a structural component of O-linked glycoproteins in mucosal membranes (Kitaoka, 2012). In several *Bifidobacterium* species, a GNB/LNB pathway is used for this particular metabolization (figure 1.2), and this consists of several different components, where the enzyme GNB/LNB phosphorylase (GLNBP, EC 2.4.1.211) is central. (Kitaoka, 2012). GLNBP hydrolyze the bond between the two LNB components Gal and GlcNAc (GalNAc in GNB). Gal1P, generated from GNB and LNB, has to be converted to Glc1P to further be able to attend energy obtaining pathways such as the bifid shunt. In the study done by (Kitaoka, 2012), GLNBP was found in all species that commonly are found in infants, such as *B. longum* subsp. *infantis*, *B. longum* subsp. *longum*, *B. bifidum* and *B. breve*. In contrast, the enzyme was not found in two species more common in an adult microbiota: *B. adolescentis* and *B. catenulatum* (Kitaoka, 2012).

The GNB/LNB pathway is a way for *Bifidobacterium* to break down galactose and is a more energy-saving variant of the Leloir pathway, which is a known galactose utilizing pathway in several bacteria (De Bruyn et al., 2013). LNB enter the GNB/LNB pathway, but galactose alone is released by β-galactosidase from the lactose unit in HMO (figure 1.2). Results from (De Bruyn et al., 2013) suggest that galactose primary is metabolized by the Leloir pathway, together with galactose-1-phosphate (Gal1P) from the GNB/LNB pathway. In theory, to utilize LNB, the bacteria only need N-acetylhexosamine-1-kinase (NahK, EC 2.7.1.162) and GLNBP from the GNB/LNB pathway, but when utilizing GNB they need the whole enzyme package from the GNB/LNB pathway. After the action of GLNBP, the GNB/LNB pathway uses the enzymes NahK, to catalyze the reaction from N-acetylglucosamine (GlcNAc) to N-acetylglucosamine-1-phosphate (GlcNAc1P) (GalNAc to GalNAc1P in GNB). It then uses both UDP-glucose-hexose-1-phosphate uridylyl transferase (GalT2, EC 2.7.7.12) and UDP-glucose/GlcNAc 4-epimerase (GalE2, EC 5.1.3.2) to catalyze the reaction from Gal1P to glucose-1-phosphate (Glc1P) from both LNB and GNB and the reaction from N-acetylgalactosamine-1-phosphate (GalNAc1P) to GlcNA1P in GNB (Kitaoka, 2012) (figure 1.3b).

In the Leloir pathway galactose is converted to Gal1P by galactokinase (GalK, EC 2.7.1.6), and further Gal1P is converted to Glc1P by both UDP-glucose-hexose-1-phosphate uridylyl

transferase (GalT1, EC 2.7.7.12) and UDP-glucose/GlcNAc 4-epimerase (GalE1, EC 5.1.3.2) (figure 1.3a). GalT1 used in the Leloir pathway and GalT2 used in the GNB/LNB pathway have an amino sequence identity of ~12 %, and GalT1 shows a higher activity in converting Gal1P to Glc1P. GalT2 also showed more activity towards GalNAc1P than to Gal1P (De Bruyn et al., 2013). Suggested by (De Bruyn et al., 2013) *Bifidobacterium* can therefore use GalT1 and GalE1, which is part of the Leloir pathway to utilize Gal1P in LNB after GLNBP has done its job, whereas they must use GalT2 and GalE2 to utilize GNB.
Sequences coding for GalT1 and GalT2 does not usually exist in the same organism, but there are some exceptions to *Bifidobacterium* and some *Clostridiales*. Both genes *galT1* and *galT2* are found in *B. bifidum*, *B. longum* and *B. breve*. These are the same bacteria that has the GNB/LNB pathway, so the coexistence can be coupled with this phenomenon (De Bruyn et al., 2013).

When oligosaccharides are metabolized to monosaccharides by various glycosidases and further degraded, hexose sugars enter the bifid shunt (figure 1.2). This is a carbohydrate fermentative pathway found in *Bifidobacterium* species, which is centered around the enzyme fructose-6-phosphate phosphoketolase (F6PPK, EC 4.1.2.22) (Sela et al., 2008). This enzyme catalyzes the following reaction: D-fructose-6-phosphate + phosphate → acetyl phosphate + D-erythrose-4-phosphate. The bifid shunt produces 1.5 moles acetate and 1 mole lactate for every mole hexose that enters (Sela et al., 2008).

A possible L-fucose utilization pathway for *Bifidobacterium* species may impact the intestinal SCFA balance due to the fact that some *Bifidobacterium* species are able to produce 1,2-propanediol (1,2-PD) from L-fucose (figure 1.2). 1,2-PD is a precursor for intestinal propionate formation. Usually, several *Clostridia* species and *Escherichia coli* are able to transform L-fucose to 1,2-PD (Bunesova et al., 2016). The study done by (Bunesova et al., 2016) describes two different pathways used by bacteria to utilize L-fucose, where one involves phosphorylated intermediates and the other does not. They found that *B. longum* subsp. *infantis* was the infant *Bifidobacterium* species that could best degrade L-fucose and suggested the use of the non-phosphorylated pathway. This pathway yields L-lactate and pyruvate, but not 1,2-PD. 1,2-PD is thought to be produced through a modified non-phosphorylated pathway (Bunesova et al., 2016).

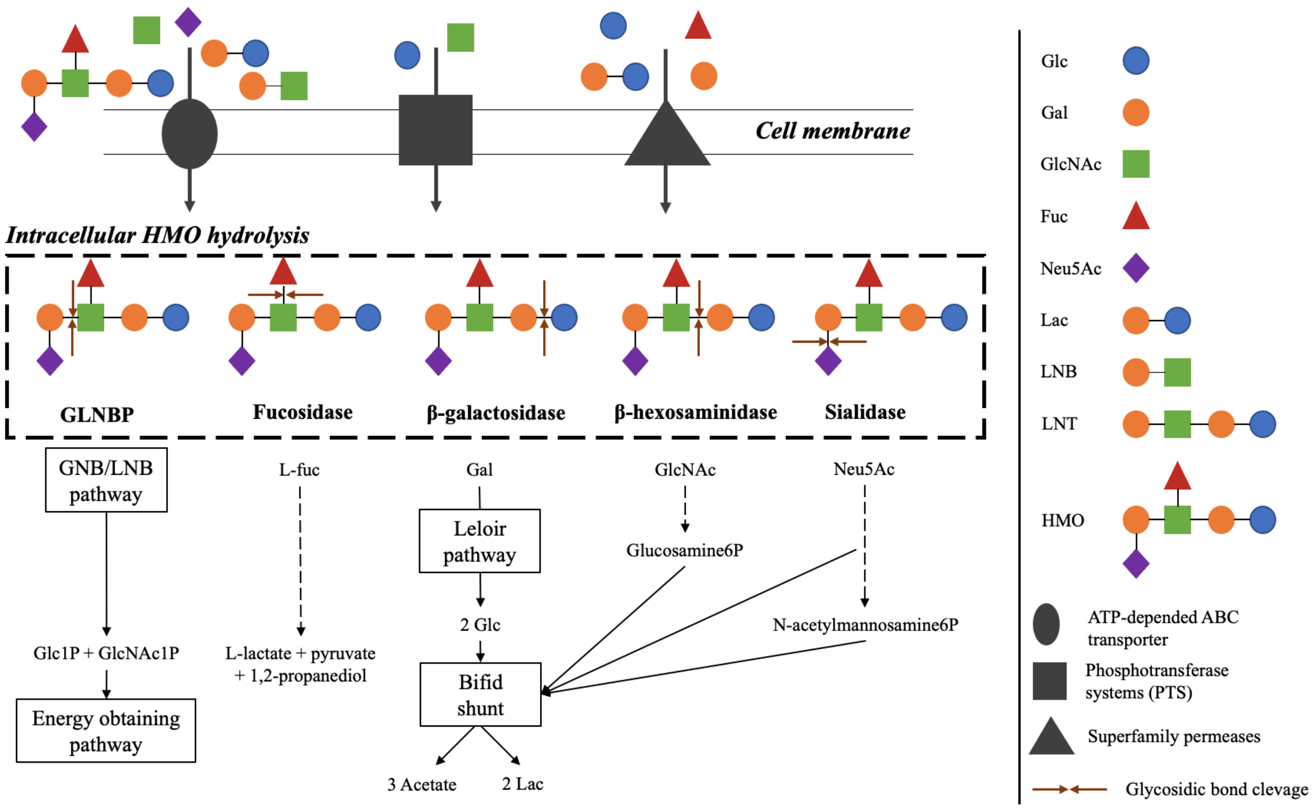


Figure 1.2: Simplified illustration of the transport and processing of HMO and derivatives. The metabolism is mainly from *B. longum* subsp. *infantis*, but several pathways and enzymes can also be found in the other *Bifidobacterium* species. HMO and its derivatives are transported over the membrane by one of three transporters, before intracellular glycosyl hydrolases process the sugars to smaller components. These components will be further degraded in one of the catabolic pathways, where the central fermentative pathway is bifid shunt. GLNBP, GNB/LNB phosphorylase; Glc, glucose; Gal, galactose; GlcNAc, N-acetylglucosamine; Fuc, fucose; Neu5Ac, N-acetyl neuraminic acid (sialic acid); Lac, lactose; LNB; lacto-N-biose; LNT; lactose-N-triose; HMO, human milk oligosaccharide; P, phosphate. The figure is modified and redrawn from (Sela et al., 2008).

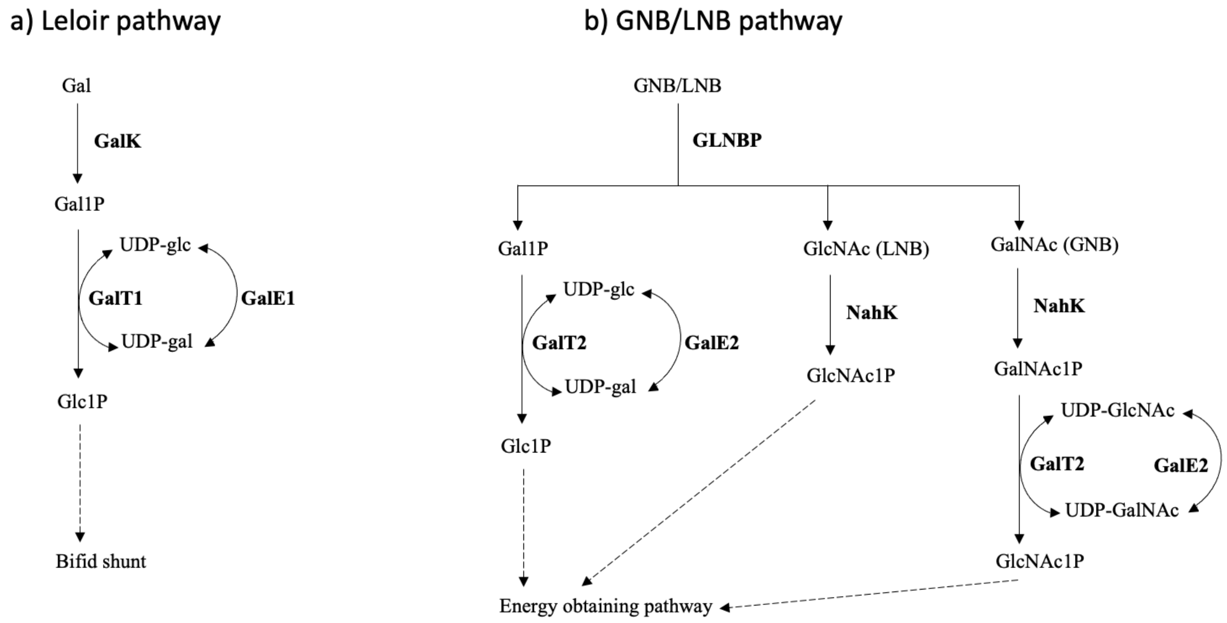


Figure 1.3: Overview over the Leloir pathway and GNB/LNB pathway. Galactose is mainly utilized in the Leloir pathway, shown in a), and some research suggest Gal1P from the GNB/LNB pathway, shown in b), also is broken down in the Leloir pathway. The GNB/LNB pathway is necessarily to utilize GalNAc1P from GNB. Gal, galactose; P, phosphate; Glc, glucose; GlcNAc, N-acetylglucosamine; GalNAc, N-acetylgalactosamine; GNB, galacto-N-biose; LNB, lacto-N-biose; GalT1/GalT2, UDP-glucose-hexose-1-phosphate uridylyl transferase; GalE1/GalE2, UDP-glucose/GlcNAc 4-epimerase; GalK, galactokinase; GLNBP, GNB/LNB phosphorylase; NahK, N-acetylhexosamine-1-kinase. The figures are made based on inspiration from (De Bruyn et al., 2013; Kitaoka, 2012).

1.5 Short chain fatty acids

The main metabolite produced from oligosaccharide degradation in the infant gut is SCFAs. In the article from (Morrison & Preston, 2016) short chain fatty acids (SCFAs) are described as "the primary end products of fermentation of non-digestible carbohydrates (NDC) that become available to the gut microbiota". SCFAs are also known as volatile fatty acids (VFAs) and consists of one to six carbons, where the most common are acetate (C2), propionate (C3) and butyrate (C4), present in the molar ratio of 60:20:20 (den Besten et al., 2013). These SCFAs have, in moderate amounts, healthy effects on the gut, and for example is butyrate the main energy source for colonocytes (Morrison & Preston, 2016). SCFA production in the gastrointestinal tract can lead to reduced pH, more accessible calcium and magnesium, and inhibition of potential pathogens (Wong et al., 2006). New studies have shown that SCFAs can be used as a signaling molecule between gut microbiota and host, and they are for

example ligands for the free fatty acid receptor 2 and 3 (FFAR 2/3). These receptors are found on immune cells and enteroendocrine cells, in addition to several other cell types (Morrison & Preston, 2016).
Fermentation of indigestible foods by *Bifidobacterium* is often linked with production of acetate (O'Callaghan & van Sinderen, 2016). There are many bacteria groups that produce acetate, but pathways for production of propionate, butyrate and lactate is more conserved and are seen in specific bacteria groups or for specific substrates. The main producers of butyrate are *Faecalibacterium prausnitzii*, *Eubacterium rectale*, *Eubacterium hallii* and *Ruminococcus bromii* (Morrison & Preston, 2016).

A biological gradient exists for each SCFA from the gut lumen to central organs. This leads to different exposure of SCFAs on different tissues and cells. The SCFAs are produced in the gut lumen, and the majority of butyrate absorption happens by the epithelium. Uptake of propionate is manly in the liver, and acetate is exposed to more of the central organs, such as muscles, the adipose tissue and the brain (Morrison & Preston, 2016), and are mainly metabolized in the liver and muscle cells (Wong et al., 2006).

1.6 Analytical methods

1.6.1 Techniques to analyze short chain fatty acid composition

About 80-90 % of the SCFAs are absorbed by the gut, and the rest will be excreted from the body (Tangerman & Nagengast, 1996). This makes it hard to analyze the amount of SCFAs produced in the intestine by just analyzing the feces, which is the most used material to analyze SCFA composition in humans, due to its easy accessibility (Primec et al., 2017). There are several different methods used to analyze SCFAs from feces, and the dominating are: gas chromatography (GC), high performance liquid chromatography (HPLC), nuclear magnetic resonance (NMR) and capillary electrophoresis (CE), where the former method is predominantly used (Primec et al., 2017).

Gas chromatography

Gas chromatography (GC) is a method used to separate and analyze organic material, by the use of a mobile and a stationary phase (Primec et al., 2017). The mobile phase is a carrier gas, that transport the sample through the stationary phase, which is the column, and into a detector. During this path, the samples will be separated based on several different factors, such as molecular weight, melting point and column temperature, and the components will be analyzed by a computer (Vitha, 2016).

The mostly used carrier gases are helium, hydrogen, argon and nitrogen. They have different properties for example in terms of separation efficiency, viscosity and speed, and must be chosen based on the column and detector used. This is because it is important that the carrier gas does not react with the stationary phase in the column (Vitha, 2016). Two different columns can be used: packed or capillary, and the most common detector used is the flame ionization detector (FID) (Primec et al., 2017). This detector breaks down organic components in the samples, which escapes the column with the carrier gas, and is mixed with hydrogen. When the organic components reach the flame, they are ionized and collected by an electrode where they produce a signal that is exported to a computer program (Vitha, 2016).

1.6.2 Sequencing methods for analyzing bacterial composition in the gut microbiota

The breakthrough for studying and classifying microorganisms came in 1977, where Carl Woese suggested using ribosomal RNA genes as molecular markers, and Fred Sanger developed the Sanger sequencing method (Sanger et al., 1977; Woese & Fox, 1977). Sanger sequencing is today known as a first-generation sequencing method. The Sanger sequencing technology is a method where a polymerase chain reaction (PCR) reaction occurs with both deoxynucleotides (dNTPs) and labeled 2´,3´-dideoxynucleotides (ddNTPs) present. When elongation takes place, some strands incorporate ddNTPs, and the elongation will be terminated. The strands, which will have different lengths dependent on when termination occurred, will be separated on a gel, and by the pattern of the bands, nucleotides could be identified, thus revealing the sequence (Sanger et al., 1977). Sanger sequencing is still used today, and with improvements it can now achieve read lengths up to ~1000 bp (Shendure & Ji, 2008). After a time of Sanger sequencing dominating the field, more companies wanted to make better sequencing technologies, and thus the second-generation sequencing, also known as next generation sequencing, was formed.

There are several different sequencing platforms belonging to next generation sequencing, but the concept of the work flow is similar between them all (Shendure & Ji, 2008). Genomic DNA is fragmented and ligated with common adapters *in vitro*. Through one of several approaches available, including *in situ* polonies, emulsion PCR and bridge PCR, millions of spatially immobilized PCR colonies are generated, where each colony has several copies of a single library fragment. Alternating cycles involving enzymatic extension reactions and imaging-based detection summarizes the sequencing process. The immobilization of colonies makes it possible to use a single reagent volume to enzymatically manipulate the array, which is a huge advantage compared to the Sanger sequencing (Shendure & Ji, 2008). A known, and much used, next generation sequencing technology is made by Illumina.

Illumina´s technology uses the sequencing by synthesis (SBS) principle, and their work flow includes four steps: library preparation, cluster generation, sequencing and data analysis (Illumina Inc., 2017). During library preparation adapters are ligated to random DNA fragments before they are amplified and purified by PCR and gel electrophoresis respectively. The library is then applied to a flow cell, where the surface is covered with surface-bound complementary sequences to the library adapters. The bound fragments will be amplified into clonal clusters through bridge PCR and now work as templates, and this completes the second step, which is cluster generation. According to (Illumina Inc., 2017) they use a "reversible terminator-based method that detects single bases as they are incorporated into DNA template strands". The dNTPs that are detected are fluorescently labeled, and the emission wavelengths and intensity during imaging of the flow cell will identify the incorporated base. The dNTPs contain a reversible terminator that blocks binding of the next dNTP. When the base has been identified, the terminator will be cleaved, and the next dNTP can bind the template. During each cycle, all dNTPs are present, compared to other technologies, which will reduce raw error rates. The last step is data analysis, where the identified sequence reads will be compared to a reference genome (Illumina Inc., 2017). Illumina has several sequencing systems for different scales. MiSeq is used for small genome and target sequencing, NexSeq is used for genome, exome and transcriptome sequencing, and HiSeq is used for production-scale genome, exome and transcriptome sequencing (Illumina Inc., 2017).

A disadvantage with the second-generation sequencing is short reads, and some companies have developed sequencing technologies with longer read length. Third-generation sequencing, also called long-read sequencing, is still a fairly new sequencing generation. There are two main types of third generation sequencing: single-molecule real-time (SMRT) sequencing and synthetic sequencing (Goodwin et al., 2016). The single-molecule approach does not create clonal clusters of amplified DNA fragments to get detectable signals, such as short-read sequencing does. Two wildly used single-molecule long-read technologies are PacBio and MinION from Oxford Nanopore Technologies (ONT) (Goodwin et al., 2016).

1.6.3 Technologies used for gene expression analysis

mRNA is the precursor to proteins and gives an indication of protein production and activity in microorganisms. mRNA degrade rapidly, and in order to analyze, it is therefore necessary in gene expression studies to convert mRNA into complementary DNA (cDNA), which is more stable. Once cDNA is made, gene expression can be analyzed by different methods. One method is RNA sequencing (RNA seq), which is a recently developed method that has taken over some other technologies, such as microarrays. RNA seq uses high-throughput sequencing methods such as Illumina (Wang et al., 2009). The sequencing steps are fairly similar to the ones described previously. Another way to analyze gene expression is through quantitative polymerase chain reaction (qPCR).

qPCR is a highly used method to measure the number of specific cDNA target copies (Costa et al., 2013). In gene expression analysis, qPCR uses PCR technology to amplify cDNA to produce high enough concentrations for fluorescence detection and quantification. The fluorescence dye is added to the samples prior to the qPCR, and during amplification they will send out signals when bound to double stranded DNA (dsDNA) (Hollister et al., 2015). Few amplification cycles (qPCR cycles) before a reached threshold value, means a greater quantity of the target material from the start. The number of PCR cycles when reached threshold value is often referred to as the Ct or Cq value (Wong & Medrano, 2005). The difference between qPCR and PCR is that in qPCR the amount of PCR products will be measured after each amplification cycle, whereas in PCR the amount of products are only measured at the end of the procedure. The PCR procedure consists of 3 steps: denaturation, annealing and elongation. During denaturation, dsDNA is parted to single stranded DNA (ssDNA) under high temperatures. The reason behind denaturation is to attach primers during annealing. The temperature rises again during elongation where dNTPs are attached to create a complementary strand to the template ssDNA (Hollister et al., 2015).

When using qPCR, you are limited to a lower number of genes, and this method can only find known sequences, based on chosen primers. qPCR is on the other hand effective for low target numbers (Illumina Inc, 2019).

1.6.4 Techniques to analyze protein composition

A huge part of protein analysis is separation. There are different techniques available for protein separation, such as gel filtration, chromatography and electrophoresis. A common separation method is the polyacrylamide gel electrophoresis (PAGE) (Lesk, 2016). PAGE involves an electric field that makes the proteins move in polyacrylamide gels. The gels are equipped with tunnels in different sizes, which makes smaller molecules travel faster. Proteins have different mobility, which depend on mass and shape, that makes them move differently through the gel. To have a separation based only on mass, proteins have to be denaturated, and a known detergent that help denature proteins are the negatively charged sodium dodecyl sulphate (SDS). When SDS-PAGE is carried out, proteins are spread out in bands, and staining with Coomassie Blue is often done to visualize these bands (Lesk, 2016).

In PAGE, complex protein mixtures can be poorly separated due to overlapping bands in the lanes. A two-dimensional PAGE is more suited to complex mixtures. They involve a two-step procedure, where proteins are first separated according to charge, then according to size. The second step occurs 90 degrees from the original direction, to create a two dimensional separation (Lesk, 2016).

Difference gel electrophoresis (DiGE) is another electrophoresis method that has the same principles as the two-dimensional PAGE, but makes it possible to compare different protein mixtures on separate gels, due to identical separation conditions for each sample (Lesk, 2016).

The separation techniques give information about some protein features, such as mass, charge and size, and the separation makes it possible to isolate the proteins and process them for further identification. To identify proteins, the most used method is mass spectrometry. This method is efficient, whilst also accurate and precise. Summarized briefly, mass spectrometry characterizes molecules by measuring their ion masses (mass/charge ratio) in a vapored stage (Lesk, 2016). The setup of a mass spectrometer consists of an ion source, a mass analyzer and a detector. The mass analyzer will measure the mass/charge ratio, and at each mass/charge ratio value, the detector will register the number of ions (Aebersold & Mann, 2003). For evaporation and ionization of the peptides there are two common methods: electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). A highly used mass spectrometry approach is the liquid chromatography-tandem mass spectrometry (LC-MS/MS). In this method, fragmented peptides are separated by liquid chromatography before they are converted into highly charged droplets by an electrospray ion source (Aebersold & Mann, 2003). When the droplets enter the mass spectrometer they are dissolved by heat, creating ions. In the first MS, specific ion masses will pass through the mass analyzer, one at a time, before they will go through a collision cell where they will be fragmented by a neutral gas. The fragmented ion will then pass through the second MS that will filter the ions based on mass/charge ratio through a second mass analyzer. The ions are then detected, and a mass spectrum is generated. The outcome of MS/MS can be used to identify the peptides (Aebersold & Mann, 2003).

1.7 The PreventADALL study

Research about the infant human gut microbiota, and how it can be connected to development of diseases later in life, are becoming a larger field of interest. One study that addresses this exact topic is the PreventADALL (Preventing Atopic Dermatitis and ALLergies) study (Lødrup Carlsen et al., 2018). This study aims to provide more information about how to prevent atopic dermatitis and allergies in infants and children. Through several years, they collected samples and information from mother-child pairs, where 2386 mothers participated, and in total 2397 children were born. According to (Lødrup Carlsen et al., 2018) all infants were randomly selected into 4 groups at birth, where "(1) no intervention; (2) skin care (oil-bath at least 5 days per week from 0.5 to 9 months of age); (3) consecutive introduction, between 3 and 4 months of age, of peanut, milk, wheat and egg at least 4 days per week complementary to breastfeeding; or (4) both interventions". Both biological samples and questionnaires were collected, and amongst the biological samples, fecal samples were collected. This was from mothers at 18 weeks pregnant, infants as newborn (meconium), infants at 3, 6, 12 and 36 months.

1.8 Aim of thesis

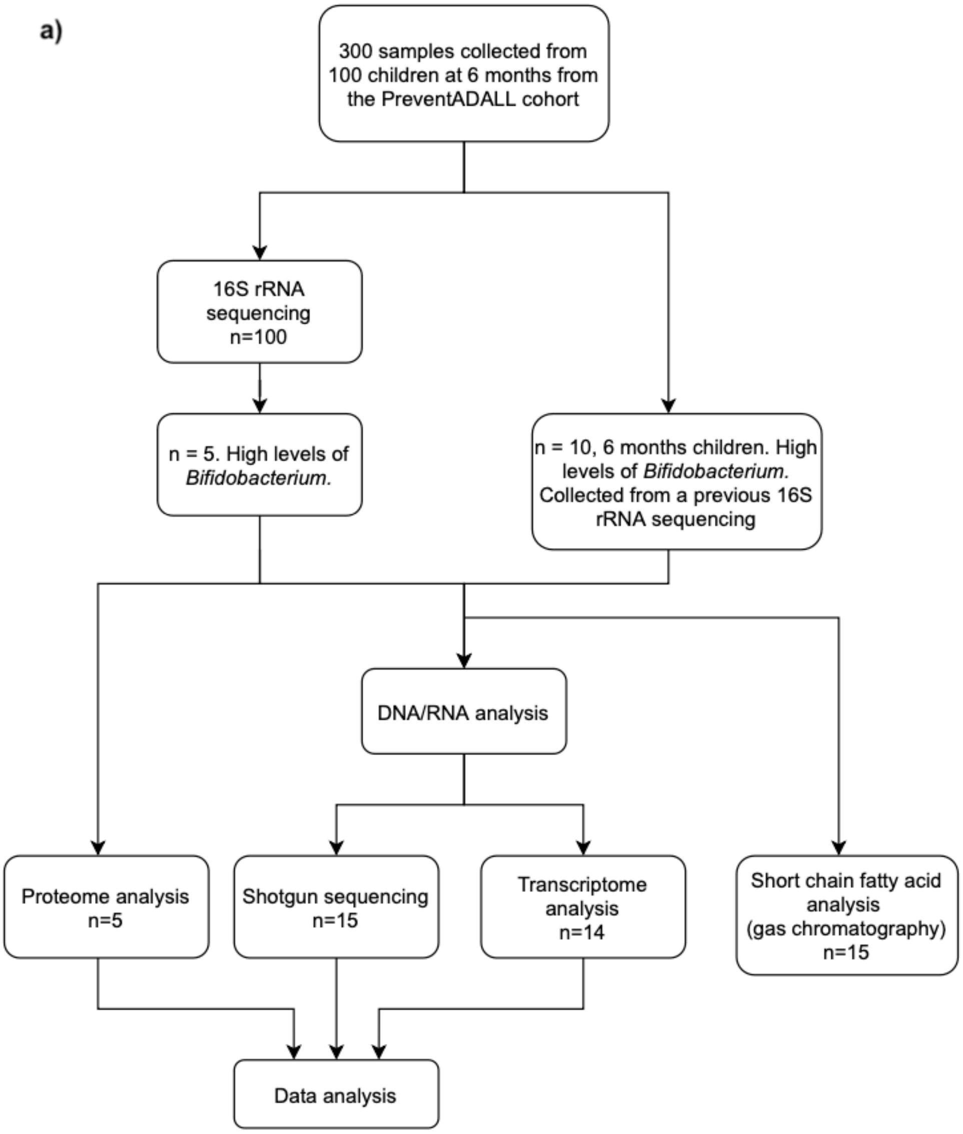
There is currently a knowledge gap related to how *Bifidobacterium* utilize HMOs *in vivo* in the infant gut. Most of the current knowledge is based on *in vitro* studies and animal studies, that has given indication that most of the *Bifidobacterium* genus express proteins that are involved in HMO utilization. To gain a deeper understanding on this topic, we have to study protein expression. To my knowledge there has not been done any proteome analyzes on *Bifidobacterium* species from the infant gut.

The aim of this study is to analyze how *Bifidobacterium* degrade HMOs in the infant gut through a multiomic approach, using fecal samples from the PreventADALL study. To achieve this, the following subgoals were included:

- find proteins that are involved in HMO utilization with use of proteome analysis
- examine short chain fatty acid composition from *Bifidobacterium*-rich samples with use of gas chromatography
- identify mRNA that can be linked to HMO utilizing proteins with use of qPCR analysis

2 Materials and methods

An overview of the experimental procedures, and the division of samples into datasets, used in this master´s thesis is illustrated in figure 2.1.


a)
300 samples collected from 100 children at 6 months from the PreventADALL cohort
16S rRNA sequencing n=100
n = 5. High levels of *Bifidobacterium.*
n = 10, 6 months children. High levels of *Bifidobacterium.* Collected from a previous 16S rRNA sequencing
DNA/RNA analysis
Proteome analysis n=5
Shotgun sequencing n=15
Transcriptome analysis n=14
Short chain fatty acid analysis (gas chromatography) n=15
Data analysis

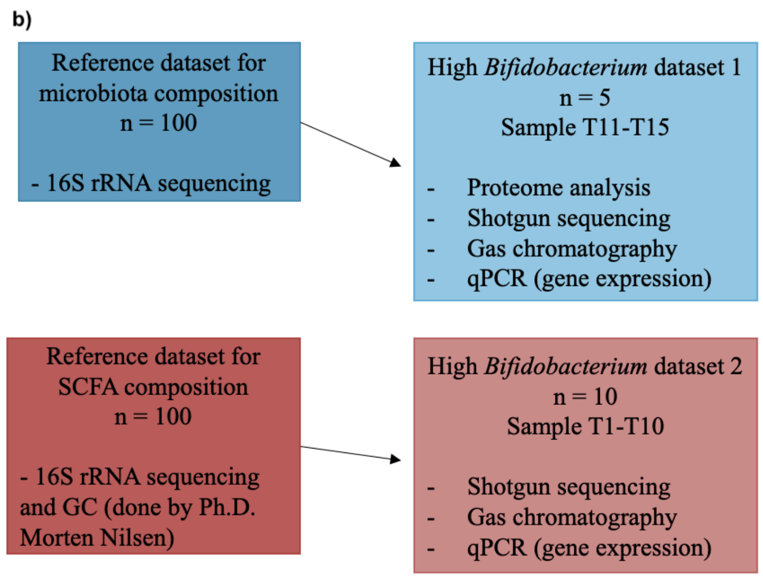


Figure 2.1: Flow chart showing the experimental procedures, and an overview of the division of datasets, in this thesis. a) shows the workflow of the thesis, and b) shows the different datasets with belonging samples and analyzes performed. Fecal samples were collected from the PreventADALL cohort. 10 samples with high *Bifidobacterium* levels were chosen from a previous 16S rRNA sequencing, done by Ph.D. Morten Nilsen. Due to a lack of samples for protein analysis, a 16S rRNA sequencing was preformed to collect 5 more samples with high levels of *Bifidobacterium*. All 15 children were breastfed at 6 months. Shotgun sequencing was performed to determine the composition of *Bifidobacterium* species, and get an overview of the genome (n=15). Transcriptome analysis, by qPCR quantification of gene expression, was done to check for potensial of HMO utilization (n=14), and proteome analysis was done to find HMO utilizing proteins in the bacterial cells and construct potential pathways (n=5). Short chain fatty acid (SCFA) composition was determined by the use of gas chromatography (n=15). Data analysis from the 16S rRNA sequencing data, preparation of shotgun data and preparation of proteome data from LC-MS/MS were done by Ph.D. Morten Nilsen.

2.1 Sample selection and preparation

Samples used in this experiment were feces samples from 6 months old children, obtained from the PreventADALL cohort (Lødrup Carlsen et al., 2018). These samples were stored in three parallels. Samples used for DNA analysis were diluted 10x with stool DNA stabilizer (PSP Spin Stool DNA Plus Kit, Invitek Molecular), samples used for RNA analysis were diluted 10x with RNA/DNA shield, and samples for protein analysis were stored without buffer. All samples were originally stored at -80 °C.

2.2 Nucleic acid based methods to analyze gut microbiota and gene expression

2.2.1 DNA/RNA extraction and purification

Mechanical lysis

Mechanical lysis was performed by adding 200 µL supernatant from pre-centrifuged 1 mL feces sample to FastPrep tubes (MP Biomedicals, USA) with 0.2 g acid-washed glass beads (Sigma-Aldrich, Germany, <106 µm), 0.2 g acid-washed glass beads (Sigma-Aldrich, Germany, 425-600 µm) and 2 acid-washed beads (2.5-3.5 mm, Sigma-Aldrich, Germany). To disrupt the cell wall, independent of cell type with the glass beads, the samples were processed in FastPrep 96 (MP Biomedicals, USA) twice at 1800 rpm for 40 sec. Then the samples were centrifuged at 13000 rpm for 5 min to collect the cell remains, such as membranes, proteins, salts and other large particles in a pellet.

In order to dissolve the remaining intact cell membranes, lysis buffer was added to the samples, and Proteinase K (ThermoFisher Scientific, USA) was added to degrade proteins that potentially could interfere with DNA, for example nucleases. The chemical lysis was done both manually and automatically using the King Fisher Flex robot (Thermo scientific, USA).

In addition to lysis buffer, 100% isopropanol was used before RNA extraction in order to release RNA from proteins in the cell, and therefore increase binding of RNA to beads during extraction.

DNA extraction

DNA was extracted both automatically, using the KingFisher Flex robot (Thermo scientific, USA), and manually. In order to achieve pure DNA, silica coated paramagnetic particles (Mag particles) from the MagMidi LGC kit (LGC Genomics, UK) were used, to selectively bind DNA to the silica surface and create a salt bridge in the presence of high salt concentrations. These salt concentrations were made by adding buffers, containing alcohol and salts, that also made it possible to cleanse the samples from impurities. An elution buffer was then added to release DNA from the silica particles, by interrupting the bridge between DNA and the surface of the silica particles, leaving the DNA in the solution.

RNA extraction and cDNA synthesis

RNA was extracted manually using MagMax 96 total RNA isolation kit (ThermoFisher Scientific, USA). The kit contained RNA binding beads and a buffer that enhanced the binding process, wash buffers that remove impurities and DNase. DNase was added to enzymatically degrade residual DNA in the solution but is not sufficient to remove all DNA. An additional DNase treatment was therefore added after elution, which was done in small volumes to concentrate the samples.

The additional TURBO DNA-free procedure was done following the manufacturer recommendation, and the routine DNase treatment, using the TURBO DNA-free kit (ThermoFisher Scientific, USA). This treatment will leave any DNA residues in the pellet and RNA in the supernatant.

cDNA synthesis was performed by combining the following reagents to 5 µL template RNA: 1x RT Reaction Premix with Random Primers (Solis BioDyne, Estonia) and 1.5 µL FIREScript Enzyme Mix (Solis BioDyne, Estonia) with a total volume of 20 µL. To control the amount of DNA left in the samples, one parallel of each sample was synthesized without FIREScript Enzyme Mix.

cDNA was synthesized using the following PCR-program: primer annealing at 25 °C in 10 min, reverse transcription at 50 °C in 60 min for maximum yield, enzyme inactivation at 85 °C in 5 min and 10 °C in ∞.

2.2.2 Nucleic acid quantification and quality control

qPCR targeting 16S rRNA V3-V4 region

The qPCR reaction mix consisted of: 1x HOT FIREPol EvaGreen qPCR supermix (Solis BioDyne, Estonia), 0.2 µM Forward primer and Reverse primer, 2 µL extracted DNA. The volume in total was 20 µL. Following program was used to amplify DNA on LightCycler 480 (BioRad, USA): 95 °C in 15 min and 40 cycles of 95 °C in 30 sec, 55 °C in 30 sec and 72 °C in 45 sec. For 16S rRNA sequencing, the following primers were used: (341F) and (806R) (see table 2.1 below).

For transcriptome analysis, 2 µL cDNA was combined with 10 µM of several different primers listed in table 2.1. The following qPCR program was used: 95 °C in 15 min and 40 cycles of 95 °C in 30 sec, 60 °C in 30 sec and 72 °C in 45 sec.

Table 2.1: Primes used to check gene expression of specific HMO-associated *Bifidobacterium* genes. The primes below came out best from a test we did with multiple primer candidates. The primers 341F and 806R were used as control.

Primers	**Sequence**	**Gene coding proteins**	**Reference**
Blon_2334F	5´- CATCACCGAGCAGGACATGA	β-1,4-galactosidase	(Yoshida et al., 2011)
Blon_2334R	5´- GCCGTACTCGTCGCACAGT		
Blon_2335F	5´- CCTGTTCAACCAGGATGAGTC	1,2-α-L-fucosidase	(Sela et al., 2012)
Blon_2335R	5´- CCGTCCACGACGAAGTAG		
Blon_2336F	5´- ATCACGCTCACCCTCCC	1,3/4-α-L-fucosidase	(Sela et al., 2012)
Blon_2336R	5´- ACATCGTCGAAGCGGAGT		
Blon_2348-2F	5´- TGGCCGTGTGATGCTGAA	2,3/6-α-sialidase	(Sela et al., 2011)
Blon_2348-2R	5´- CCGGGAGATGGCGACATA		
Blon_2355F	5´- ACGCGCCGCGCAATAGGAAT	β-*N*-acetyl-glucosaminidase	(Garrido et al., 2012)
Blon_2355R	5´- GGACGTGACTCGTGGCCGTG		
Blon_2016F	5´- GGACCACCTTGACTTGGACAA	LNT β-1,3-galactosidase	(Yoshida et al., 2011)
Blon_2016R	5´- GTCCACTTATCTGCCTTGAAGGA		
Blon_0732F	5´- ACGCTGGACCGCACATTGGG	β-*N*-acetyl-glucosaminidase	(Garrido et al., 2012)
Blon_0732R	5´- AACGCCAGCAGTTCCTCGCC		
341F	5’- CCTACGGGRBGCASCAG		(Yu et al., 2005)
806R	5’- GGACTACYVGGGTATCTAAT		

Agarose gel electrophoresis

Either 1.5% or 2% agarose gel, consisting of agarose (Invitrogen, USA) and 1x tris-acetate EDTA (TAE) buffer, with added PeqGreen dye (Peqlab, Germany), were made. The 1.5% agarose gel was set to 80V in 30 min, and the 2% agarose gel was set to 80V in 45 min. Five microliter sample with 1x purple loading dye (New England BioLabs, USA) were applied to the gel. A 100 bp ladder (Solis BioDyne, Estonia), sometimes together with a 1 kb ladder, was used as a reference. The gel-results were visualized by UV-lights, using the Molecular Imager Gel Doc^{TM} XR Imaging System (BioRad, USA).

Measurement of DNA/RNA quantity by Qubit

Quantity of nucleic acids were measured using a Qubit Fluorometer (Invitrogen, USA). The Quant-iT^{TM} Assays Abbreviated Protocol (Invitrogen Corporation, 2007) was followed and the Quant-iT^{TM} kit (Invitrogen Corporation, USA) was used to detect quantity of nucleic acids in the samples. Used 2 µL sample to 198 µL Quant-iT^{TM} Working Solution. The Quant-it reagent contains, according to (Thermo Fisher Scientific Inc, 2018), “target-selective dyes that emit fluorescence when bound to DNA, RNA or protein”, dependent on the kit been used.

2.2.3 PCR amplification and purification

Amplification of qPCR products

To amplify template DNA from qPCR, 5 µL product was mixed with the following components: 1x HOT FIREPol Blend Master Mix Ready to Load (Solia BioDyne, Estonia), 0.2 µM Forward primer and reverse primer (Yu et al., 2005). The total volume was 25 µL. The PCR products were amplified using the following program: 95 °C in 15 min, 30 cycles of 95 °C in 30 sec, 55 °C in 30 sec and 72 °C in 45 sec, followed by 72 °C in 7 min and 10 °C in ∞. Both 5 µL template DNA and 30 cycles were to increase the DNA amount, because of low Cq-values from qPCR targeting 16S rRNA V3-V4 region. The products were checked on gel electrophoresis with 100 bp ladder (Soils BioDyne, Estonia).

Clean-up of PCR product

The clean-up of products after amplicon PCR was done automatically on Biomek 3000 (Beckman Coulter, USA). Used 1x volume of Sera-Mag beads to 10 µL PCR product and followed the manufacturer recommendation to the Biomek robot to clean up 16S samples. Some of the purified products was checked on gel electrophoresis to ensure that no product was removed during the clean-up.

2.2.4 Amplicon (16S) sequencing

Index PCR for Illumina sequencing

Purified PCR products were used as templates for the sequencing. Indexes were attached to the products, to make them separable during 16S sequencing. The index application was done using the Eppendorf epMotion 5070 robot (Eppendorf AG, Germany), with 0.2 µM concentration of each forward and reverse primer. The index primers used were F1-16 and R26-32 (supplement, table S.6) to achieve a unique combination for each sample. 1x FIREPol Master Mix Ready to Load (Solis BioDyne, Estonia) and 2 µL template DNA were then applied to the indexes, to achieve a final volume of 25 µL.

The DNA fragments were then amplified using the following PCR program: 95 °C in 5 min, 10 cycles of 95 °C in 30 sec, 55 °C in 1 min and 72 °C in 45 sec, followed by 72 °C in 7 min and 10 °C in ∞, and then checked on 1.5% agarose gel.

Quantification and Normalization

Amounts of DNA from indexed PCR products were measured using the Cambrex-FLEX 800 CSE robot (ThermoFisher Scientific, USA) to prepare for the 16S rRNA sequencing. A volume of 70 µL Quant-iT Working Solution, same solution used for Qubit measurements, was mixed with 2 µL DNA sample, and Nunc 96 Nontreated Black Microwell plates were used to measure DNA amount.

A selection of 20 samples, ranging from low to high fluorescence value, were afterwards measured with Qubit to get the concentration for making a standard curve. This standard curve was used to calculate ng/µL concentration of the other samples.

To normalize, the samples were calculated based on the sample with highest concentration. All samples with a value over 10 µL were set to 10 µL, based on the requirements of the robot used in normalization and pooling, Biomek 3000 (Beckman Coulter, USA). The pooled sample was measured by Qubit afterwards.

Clean-up of pooled library

Clean-up of pooled library with 16S products was done manually, using 1.5x volume of 0.1% Sera-Mag beads to 300 µL pooled sample. Followed the AMPure protocol and eluted in 40 µL PCR-water. The product was checked with Qubit and gel and quantified with qPCR.

KAPA Library Quantification

The KAPA Library Quantification kit for Illumina platforms (KK4828, Kapa Biosystems) was used to quantify amplicons in the pooled sample. A dilution series from 10^{-4} to 10^{-7} was made from the pooled sample, and together with 6 standards, they were quantified in duplicates. Standards, negative control and 2 µL sample were each mixed with 12 µL of a PCR mix, containing 2x KAPA SYBR FAST qPCR master mix and 10X Primer premix, and 6 µL PCR water. A qPCR was preformed, using the following cycling protocol: 95 °C in 5 min, 95 °C in 30 sec and 60 °C in 45 sec. The melt curve analysis ranged from 65-95 °C.

Used the KAPA Library Quantification Data Analysis Template to quantify the data, and to calculate back to the concentration of the pooled sample.

16S rRNA amplicon sequencing

The 16S rRNA amplicon sequencing was done using Illumina MiSeq. From the qPCR of pooled sample using KAPA Library Quantification Data Analysis Template, the pooled sample was diluted to 4 nM, using nuclease-free water. Following the protocol from Illumina MiSeq, the pooled sample was further diluted to 6 pM and combined with a PhiX control, which constituted 15% of the sample, and then applied to the Illumina MiSeq (Illumina, USA). PhiX was added to avoid cross-signals between different samples during the sequencing and is an adapter-ligated library.

Data analysis from 16S rRNA sequencing in QIIME

Sequencing data from the 16S analysis was processed by PhD Morten Nilsen, with use of the Quantitative Insights Into Microbial Ecology (QIIME) pipeline. The pipeline assembled forward and reverse reads and sorted them to their respective samples. To check reads for chimeras, Usearch was used, and the SILVA database was then used to create OTUs with ≥ 97% 16S rRNA identity and assigning taxonomy (Nilsen et al., 2020). The cut-off was set to 5000 sequences per sample.

2.2.5 Shotgun sequencing

DNA tagmentation

To fragment and tag the extracted, genomic DNA with adapter sequences, Bead-Linked Transposomes (BLT), from the Illumina DNA prep kit, were used. Thirty µL cleansed DNA was transferred to a PCR plate and combined with Tagmentation Master Mix, before the plate was tagmented during the following program on the thermal cycler (Applied Biosystems, USA): 55 °C in 15 min and 10°C in ∞ with a reaction volume to 50 µL and preheat lid option at 100 °C.

Tagmentation was stopped with Tagment Stop buffer and heat treatment with the following program: 37 °C in 15 min, and 10 °C in ∞, with 60 µL reaction volume and preheat lid option at 100 °C. The adapter-tagged DNA was then washed with Tagment Wash buffer before further processing.

Amplification of tagmented DNA

To recognize the DNA sequences after Illumina sequencing, the tagmented samples have to contain index adapters with a specific combination attached to each sample. The indexes were 24 plex individual tubes from the Illumina prep DNA kit.

The following PCR program was used: 68 °C in 3 min, 98 °C in 3 min, X cycles of: 98 °C in 45 sec, 62 °C in 30 sec and 68 °C in 2 min, followed by 68 °C in 1 min and 10 °C in ∞. The reaction volume was 50 µL and preheat lid option was set to 100 °C.

The amount of PCR cycles was calculated from the Qubit results from DNA extraction, by multiplying the result with 30 µL. This was the amount of sample applied to the PCR plate during tagmentation. In the protocol (Illumina, 2020b) a table with amount of total DNA input (ng) and corresponding number of PCR cycles are shown.

The samples were run with a PCR program with 12, 8 and 6 cycles depending on total DNA input (ng) (Supplement, table S.1).

Purification of amplified DNA tagmentations

To clean up amplified DNA tagmentations, the Library Prep Protocol from Illumina (Illumina, 2020a) was followed, with use of the Illumina DNA prep kit. Followed the clean-up method for small PCR fragments (<500bp), due to results from the gel electrophoresis, and if the method for over 500 bp was used, a lot of sample would be lost. Due to this method, the transferred sample volumes were multiplied with 1.8x to find the fitting amount of sample purification beads to add. The samples were washed in 80% ethanol.

Pooling of library

To pool the shotgun library, the method for DNA inputs less than 100 ng, from the protocol (Illumina, 2020b), was used. The samples were quantified based on the Qubit results and calculated and quantified based on the sample with highest concentration. They were then quantified again with an equal factor to reach a volume between 60 µL and 80 µL. In those cases where the concentration of the samples was too low, speedvac was used to increase the concentration. The pooled library was sequenced by Norwegian Sequencing Centre (NSC) on NovaSeq SP. The library got ½ flow cell, and the sequencing resulted in 150 bp paired end reads.

Data analysis from shotgun sequencing

The quality of the reads was checked by FastQC. Data from the sequencing was processed by PhD Morten Nilsen. Firstly, the reads were filtered and trimmed by trimmomatic, with the parameters MAXINFO: 50:0.24, LEADING: 10, TRAILING: 10, SLIDINGWINDOW: 5:20, MINLEN: 32. In other words, the reads were balanced by a read length of 50 and error rate 0.24 to maximize the value of each read. Then bases of the start and end of the read were cut if the quality was below 10. The read will then be cut if the average quality within a group of 5 bases is below the threshold set to 20. Lastly all reads below the length of 32 was removed. After trimmomatic, Bowtie2 and Samtools were used to remove human DNA sequences, and MetaSPADES was used to assemble the reads. To create bins, two separate programs were used, MaxBin and Metabat2. From these programs, the best candidates were selected with use of the program Drep. Taxonomy within each bin was performed by the Kraken2 standard Plus database, and Prodigal was used to create the amino acid sequences corresponding to each sequence in the bins.

With bins with genomic information and estimated amino acid sequences I processed the data in RStudio version 1.3.1093 and made a FASTA file with amino acid sequences only belonging to *Bifidobacterium* species. The procedure is attached as an R Markdown file in appendix E.

The FASTA file with *Bifidobacterium* species were checked in the KEGG database to see potential proteins and pathways.

The FASTA file was processed further by Prof. Knut Rudi to attach proteins to the amino acids. The different amino acid sequences were mapped to proteins by CLC Genomic Workbench and taxonomy, GO names, enzyme codes etc. were imported from InterProScan.

2.3 Protein based methods

2.3.1 Isolation of bacterial cells

Approximately 0.2g fecal sample was suspended in 10 mL ice-cold TBS-buffer in 50 mL tubes. To remove large materials and intact human cells from the samples, they were passed through a 20 µm filter, using Merck^{TM} Nylon-Net Steriflip^{TM} Vacuum Filter Unit (Fisher Scientific, USA). Centrifugation at 1500 g for 5 min can also be used for this step, to collect large particles in the pellet. The samples were then centrifuged at 4000 rpm for 10 min, to collect bacterial cells in the pellet, that was further resuspended in 10 mL cold TBS-buffer. To remove eukaryote proteins, the samples were passed through a second filter, a 0.22 µm nitrocellulose membrane filter (Millipore, USA). The bacterial cells will be captured on the filter, and eukaryotic proteins will pass through. The filtration was performed on a Millipore Vacuum Filtration System (Merck Millipore, USA).

2.3.2 Cell lysis

Filters from the isolation of bacterial cells step were cut in small pieces and placed in their respective tubes, together with 0.2 g acid-washed glass beads (Sigma-Aldrich, Germany, <106 µm), 0.2 g acid-washed glass beads (Sigma-Aldrich, Germany, 425-600 µm) and 2 acid-washed beads (Sigma-Aldrich, Germany, 2.5-3.5 mm), and 1 mL lysis buffer with 2% SDS, to perform a chemical and mechanical lysis combined. The lysis buffer worked on the cells for 30 min on ice with occasional mixing to dissolve the cell membrane, so that the SDS get access to the proteins and unfold them, before the cell wall was disrupted by 3 x 60 sec pulses on FastPrep 96 (MP Biomedicals, USA) at 1800 rpm. The samples were then centrifuged at 16000 x g for 15 min at 4 °C to collect the glass beads at the bottom of the tubes. Approximately 700 µL supernatant was transferred to new tubes.

2.3.3 Measurement of protein concentration

To measure the protein concentration, a BCA (Bicinchoninic Acid) Protein Assay was performed. One milliliter BCA working solution, consisting of 50 parts BCA and 1 part reagent from the Pierce BCA Protein Assay Kit (ThermoFisher Scientific, USA), was added to 50 µL 1/5 diluted lysed sample. The reagent in BCA working solution contains Cu^{2+}, and in order to make the proteins reduce Cu^{2+} to Cu^{+} in alkalic environments, provided by BCA, the samples were incubated at 60 °C for 30 min, then cooled down to room temperature. This will make the samples purple, and the color can be measured with absorbance at 562 nm on the Eppendorf BioPhotometer D30 (Eppendorf AG, Germany). The instrument will estimate a protein concentration for the samples based on this absorbance. Before measurement, the instrument was blanked with a negative control, containing lysis buffer, with the same treatment as the samples. The instrument was already calibrated with BCA standard solutions (25, 50, 100, 150, 200 and 250 µg/mL), that were prepared in the same way as the samples.

2.3.4 Protein purification through SDS-PAGE

Based on the experience of Magnus Arntzen, 40 µg protein in 19.5 µL sample on the SDS-PAGE (Sodium Dodecyl Sulphate – Polyacrylamide Gel Electrophoresis) gives best results on the mass spectroscopy performed later on. The lysed samples were speedvaced to achieve the desired concentration, based on the concentrations from BCA Protein Assay (supplement, table S.2). Nineteen point five microliter sample was mixed with a reducing sampling buffer, resulting in a mix consisting of 40 µg protein, 1x Sampling buffer (ThermoFisher Scientific, USA) and 1x Reducing agent (ThermoFisher Scientific, USA). The Sampling buffer gives color to the samples and make them visible in the gel. The Reducing agent consists of DTT that is known to reduce disulfide bonds in proteins, and therefore keep the proteins unfolded together with SDS that already is in the samples, when the samples are denatured at 90 °C for 5 min. After denaturation, the samples were centrifuged for 1 min at 10000 x g.

To the wells in the SDS gel (Mini-PROTEAN TGX stain-free gel, Bio-Rad Laboratories, USA), 30 µL sample was applied with blanks in between to inhibit one sample well contaminating the neighboring well, or the bands to blend into each other. The inner chamber was filled with freshly made 1x TGS-buffer (Tris-Glycine-SDS, Bio-Rad, USA), and the rest of the container with 1x used 1x TGS-buffer. The gel was set at 270V for 6 min, until the band had traveled 1 cm on the gel. SDS-PAGE was not used in this experiment as a protein separation step, but rather as a clean-up step to get as much pure protein as possible.

Staining and destaining SDS-gel

Before staining, the gel was rinsed with Milli-Q water. To make the protein bands visible, the gel was stained with a staining stock (0.05 % Coomassie Brilliant blue R-250 (Bio-Rad, USA), 25 % isopropanol and 10 % acetic acid glacial) that binds to proteins. After 1 hour staining at 20 rpm, destaining solution (staining solution without Coomassie Brilliant blue R-250) was applied to remove the blue color from the gel, in order to get visible blue protein bands. Destaining was done 2 x 20 min at 20 rpm, before an overnight destaining was performed with 1:2 dilution of the destaining solution.

2.3.5 In gel reduction, alkylation and digestion

The gel was rinsed with Milli-Q water before the bands were cut in 1x1 mm cubes and placed in their respective tubes. Two hundred microliter Milli-Q water was added to cover the gel pieces, and then the samples were incubated in 15 min at room temperature on a thermo mixer (500 rpm). The fluids were removed, and 200 µL of a solution with 50% ACN (Acetonitrile, Honeywell, USA) and 25 mM AmBic (Ammonium bicarbonate, Sigma-Aldrich, USA) was added to de-color and rinse the gel pieces. The samples were again incubated in 15 min at room temperature and 500 rpm. The liquid was removed, and the two previous steps were repeated once more. To extract all fluids from the gel pieces, 100 µL 100 % ACN was added to each sample before incubation at room temperature for 5 min and 500 rpm. The liquid was removed, and the samples air-dried for 1-2 min.

Reduction and alkylation

The disulfide bonds in the samples were reduced by adding 50 µL DTT solution, consisting of 10 mM DTT (Dithiothreitol, Sigma-Aldrich, USA) and 100 mM AmBic, and incubated for 30 min at 56 °C at 500 rpm. Once the samples were cooled down, the proteins were prevented from forming disulfide bonds by adding 50 µL IAA solution (55 mM IAA (Iodoacetamide, Sigma-Aldrich, USA), 100 mM AmBic), that binds to the thiol group on cysteins. Due to IAA light sensitivity, the samples were incubated in the dark for 30 min. IAA was then removed, and 200 µL 100 % ACN was added to extract all fluids from the gel pieces. The samples were incubated for 5 min in room temperature at 500 rpm. The fluids were removed, and the samples were air-died for 1-2 min.

Digestion of proteins

To digest the proteins to peptides, 30 µL 10 ng/µL Trypsin solution (made with a Trypsin buffer consisting of 1M Ambic and 100 % ACN) was added to the gel pieces, so that the serine protease Trypsin could cleave the protein chain at a specific place. The samples were incubated on ice for 30 min, before additional trypsin buffer was added to cover the gel pieces. The samples were the incubated over night at 37 °C at 500 rpm, before the reaction was stopped by adding 40 µL 1 % TFA (Trifluoroacetic acid, VWR, USA). To get the peptides from the gel pieces and into the TFA solution, the samples were sonicated on water bath for 15 min.

2.3.6 Extract and cleanse peptides from solution using ZipTips and NanoDrop measurement

Peptides were extracted from the solution into a hydrophobic stationary phase (C_{18} material) inside ZipTips (Merck-Millipore, USA), using a C_{18} solid phase extraction method. The binding of peptides was enhanced by conditioning the C_{18} material beforehand with 100 % methanol as an organic compound, 70 % ACN/0.1 % TFA as an acidic compound and 0.1 % TFA as an ion-pairing reagent. After binding of peptides from the sample, the peptides were washed with 0.1 % TFA, before they were eluted in 20 µL 70 % ACN/0.1 % TFA. After peptides from all samples were eluted in their respective tubes, speedvac was used to dry the samples, before cleaned peptides were dissolved in 10 µL 2 % ACN/0.1 % TFA and transferred to HPLC vials (VWR, USA). One point five microliter sample was then measured on Thermo Scientific NanoDrop One Microvolume UV-Vis Spectrophotometer (A205) (ThermoFisher, USA), and the results can be found in the supplement, table S.3. The samples were further analyzed on a nanoLC-Orbitrap MS/MS system (Dionex Ultimate 3000 UHPLC, Thermo Scientific, Germany), connected to a Q-Exactive mass spectrometer (Thermo Scientific, Germany). Details about the MS procedure is found in appendix C.

2.3.7 Data analysis from mass spectroscopy

Raw files from the mass spectroscopy were analyzed by PhD Morten Nilsen with MaxQuant version 1.6.7.0, with the MaxLFQ algorithm implemented for label-free quantitative detection of proteins. Raw files were searched against both the sequence database made in RStudio and against human genome (*Homo sapiens*, 73952 sequences), the latter to remove contaminants. Detailed information about the MaxQuant procedure is found in Appendix D.

Data from MaxQuant was processed further in Perseus version 1.6.15.0. I filtered rows based on categorical columns to remove contaminants and based on text column to remove all proteins mapped to the human genome database. The data was then log2 transformed and all missing values from the label-free quantification (LFQ) intensity, which in other samples were over 19, were replaced by the value 10 to easily work with the data. The matrix was lastly annotated by columns to the database from InterProScan with taxonomy, GO names, enzyme codes etc.

2.4 Determination of short chain fatty acid (SCFA) composition

From the 10x diluted feces samples, 200 µL was diluted 1:1 with an internal standard. This standard consisted of 0.4 % formic acid (Sigma-Aldrich, Germany), to reduce pH in the samples so SCFA can easily be activated and 2 mM 2-methylvaleric acid (Sigma-Aldrich, Germany), to keep track of any displacements whilst not interfere with the results, as it does not exist as a SCFA in the human gut. The samples were centrifuged at 13000 rpm in 10 min. The supernatant was filtered through 0.2 µm filters (VWR, USA) to remove smaller particles and centrifuged at 10000 rpm in 5 min. The fluid was transferred to 300 µL GC vials (VWR, USA). The instrument used for the gas chromatography analysis was Trace 1310 with an autosampler (ThermoFisher Scientific, USA).

A standard was run between every 5 sample to detect any changes, for example displacements, in the run. This standard consisted of 0.2 % formic acid and 1 mM of the following acids: 2-methylvaleric acid, acetic acid (Sigma-Aldrich, Germany), propionic acid (Sigma-Aldrich, Germany), isobutyric acid (Sigma-Aldrich, Germany), butyric acid (Sigma-Aldrich, Germany), isovaleric acid (Sigma-Aldrich, Germany) and valeric acid (Sigma-Aldrich, Germany). Detailed information about the gas chromatograph is listed in Appendix B. The data program used to identify peaks was the Thermo Scientific^{TM} Dionex^{TM} Chromeleon^{TM} 7 Chromatography Data System Version 7.2 SR4.

2.5 Statistical analysis

Spearman correlations were used to correlate the SCFAs to both bacterial taxa and to *Bifidobacterium* species. The significant level was set to 0.05, and the analysis was performed in RStudio version 1.3.1093. From the correlation analysis, a matrix was created, based on the Spearman´s rank correlation coefficient, rho (ρ). A matrix with p-values (pairwise two-sided p-values) was also included. These two matrices were used to create correlation plots. For more details about the statistical analysis, see the R Markdown file in appendix E.

3 Results

3.1 Gut microbiota composition in 6 months children from 16S rRNA sequencing

Overall, the microbiota composition in high *Bifidobacterium* dataset 1 and 2 has an overrepresentation of *Bifidobacterium* (figure 3.1a and b), which is due to the selection of samples with high abundance of the bacteria in these two datasets. The amount of *Bifidobacterium* is 72.9 % and 67.63 % in high *Bifidobacterium* dataset 1 and 2 respectively. This is around three times as much as in the reference dataset for microbiota composition, which consists of 23.75 % *Bifidobacterium* (figure 3.1c). Besides *Bifidobacterium*, the three most abundant bacteria in the high *Bifidobacterium* dataset 1, which was used for proteome analysis and consists of sample T11-T15, are *Escherichia-Shigella* (4.97 %), *Bacteroides* (3.61 %) and *Streptococcus* (3.19 %). *Escherichia-Shigella* is also highly abundant in the high *Bifidobacterium* dataset 2, with 7.18 %. In addition, the latter dataset, which was not used for proteome analysis and consists of sample T1-T10, has *Clostridium sensu stricto 1* (3.38 %) and *Veillonella* (3.35 %) at the top four most abundant bacteria. The data in the reference dataset for microbiota composition, represents a normal gut microbiota in 6 months old children. This microbiota is still dominated by *Bifidobacterium*, with *Bacteroides* (13.75 %), *Escherichia-Shigella* (10.54 %) and *Clostridium sensu stricto 1* (8.42 %) highly abundant.

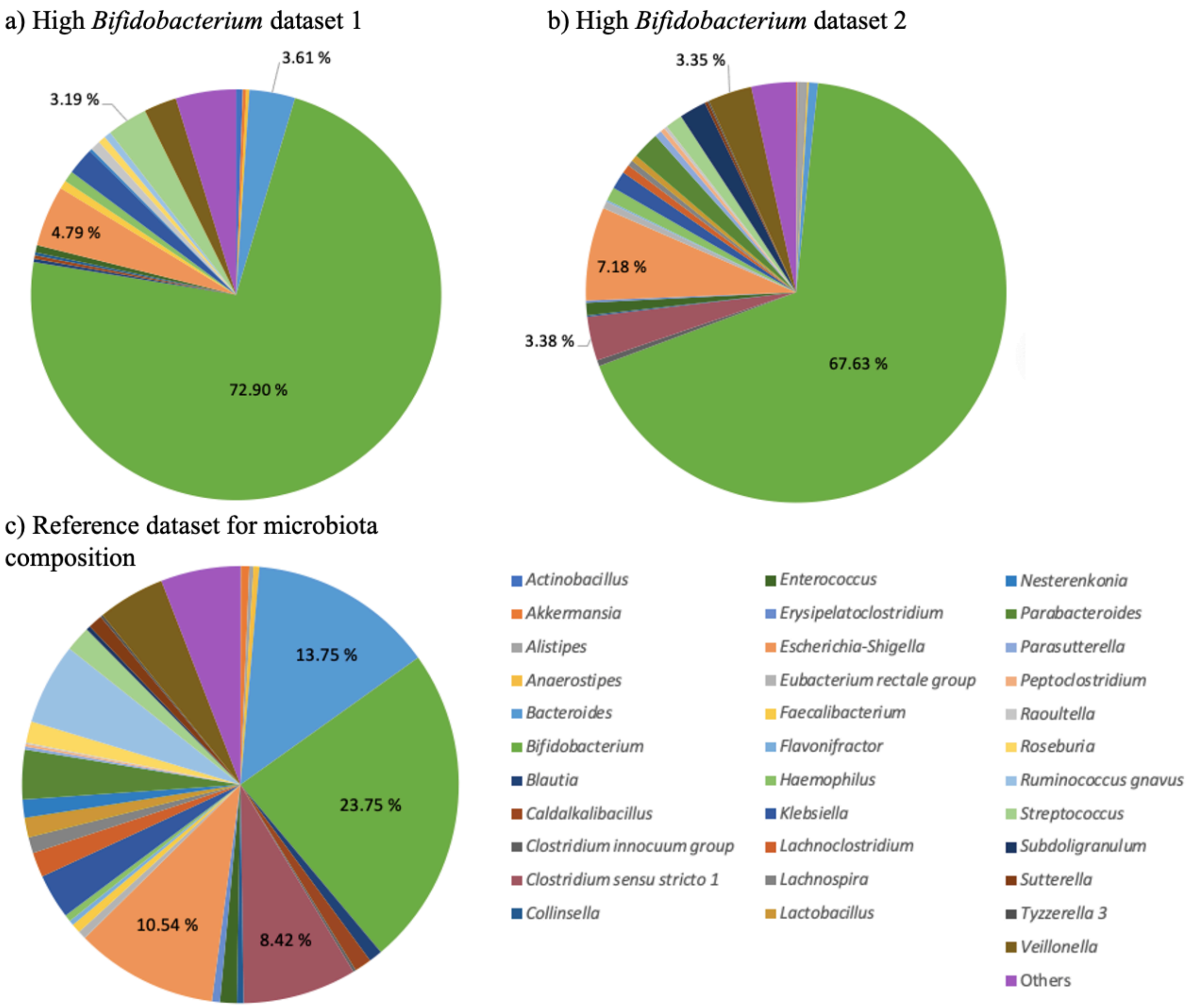


Figure 3.1: Distribution of average bacteria present over 1 %. The pie charts are based on data from the 16S rRNA sequencing. a) shows the average of bacteria in the high *Bifidobacterium* dataset 1, and b) shows the average of bacteria in the high *Bifidobacterium* dataset 2, which was sequenced by Ph.D. Morten Nilsen. c) shows the reference dataset for microbiota composition and represents the total average of bacteria in the gut of 100 children at the age of 6 months. In all pie charts, the four most dominating samples, with the exception of "others", are shown in percent.

3.2 Composition of *Bifidobacterium* species from shotgun sequencing

Based on data from the shotgun sequencing, the composition of *Bifidobacterium* species was determined in each sample and illustrated in figure 3.2. The composition of different *Bifidobacterium* species is very different in each sample, but overall, *Bifidobacterium longum* subspecies are dominating. When comparing the datasets, *B. bifidum* and *B. longum* are the dominating species in infants belonging to the high *Bifidobacterium* dataset 1, whereas *B. breve*, *B. longum* subsp. *infantis* and *B. longum* are dominating the high *Bifidobacterium* dataset 2. *B. breve* is dominating sample T1 and T2, whilst sample T3, T5, T9 and T10 are dominated by *B. longum* subsp. *infantis*. Sample T14 is completely dominated by *B. bifidum*, and this bacterium is also found in high amounts in sample T7, T11 and T12. In sample T6 there is a high abundance of *B. pseudocatenulatum*, which is not found in the other infants. *B. dentium* and *B. adolescentis* present in sample T12 and T13 respectively, are also rarely found in the other samples.

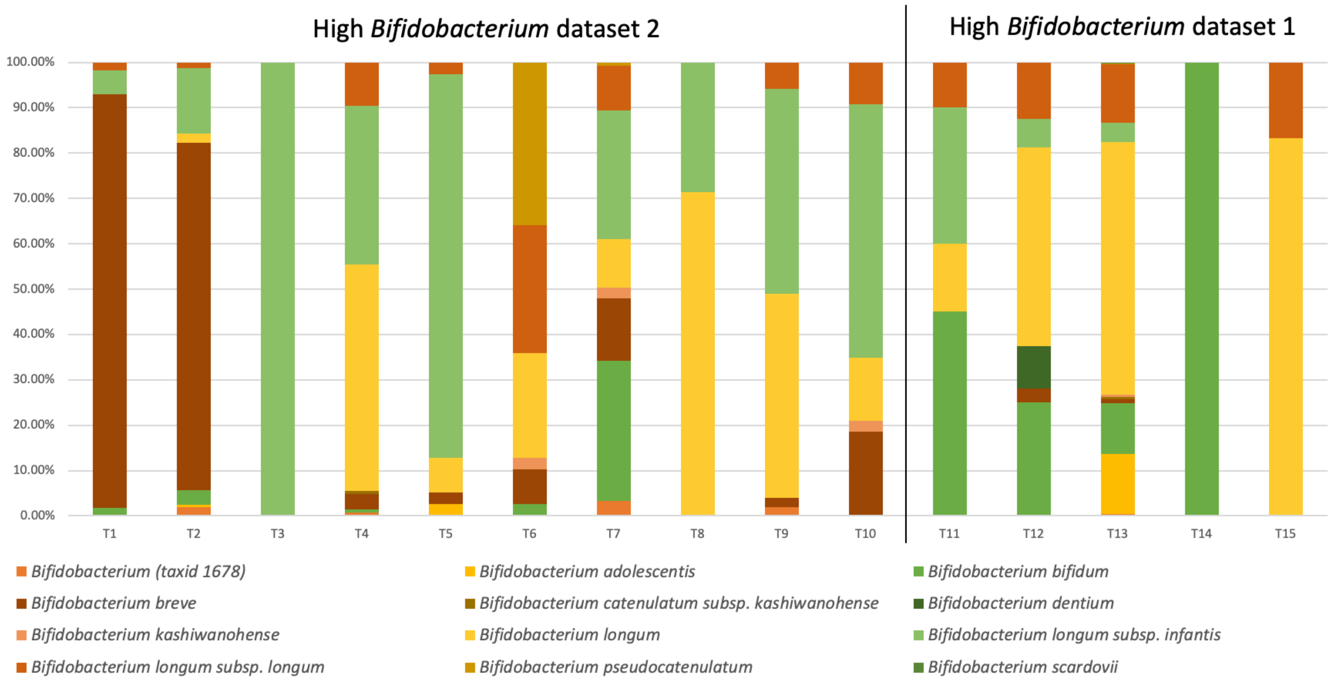


Figure 3.2: Distribution of *Bifidobacterium* species. The bar chart is based on data from the shotgun sequencing from high *Bifidobacterium* dataset 1 and 2. The species incorporated in the *Bifidobacterium longum* category are not sequences at a low enough level to be incorporated in the subspecies. This category can therefore include both *B. longum* subsp. *longum*, *B. longum* subsp. *infantis* and *B. longum* subsp. *suis*, which are the three subspecies of *B. longum*.

3.3 SCFA composition and correlation between SCFA and gut bacteria

In all three datasets presented in figure 3.3, the SCFA composition is dominated by acetic acid, with an amount of 93.7 %, 85.1 % and 87.83 % in figure a, b and c respectively. The high *Bifidobacterium* dataset 2 (figure 3.3b) resembles the reference dataset for SCFA composition (figure 3.3c) the most. The two datasets have almost equal amounts of acetic acid and propionic acid. The latter is 6.9 % in the high *Bifidobacterium* dataset 2 and 6.76 % in the reference dataset, which is almost twice as much as in the high *Bifidobacterium* dataset 1 (3.9 %) (figure 3.3a). The reference dataset for SCFA composition has twice as much butyric acid, being 4.13 %, then the two other datasets. Whilst the high *Bifidobacterium* dataset 1 only consists of three acids, the high *Bifidobacterium* dataset 2 has over the double amount of isobutyric acid (3 %) and isovaleric acid (2.3 %) than the reference dataset.

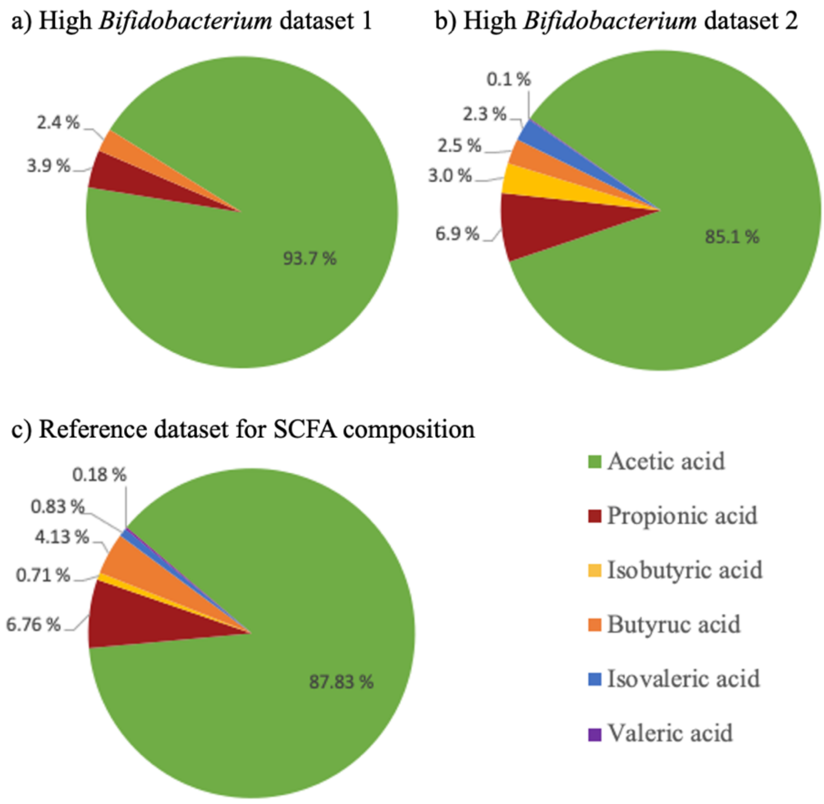


Figure 3.3: Average short chain fatty acid (SCFA) composition. The pie charts are based on data from the gas chromatography. a) shows mean SCFA composition in the high *Bifidobacterium* dataset 1 and b) shows mean SCFA composition in the high *Bifidobacterium* dataset 2. c) is used as a reference, and shows mean SCFA composition in the reference dataset for SCFA composition, which is analyzed by Ph.D. Morten Nilsen (Nilsen et al., 2020). The distribution of SCFAs in each sample from the high *Bifidobacterium* dataset 1 and 2 can be found in the supplement, figure S.1.

To check the correlation between gut bacteria and SCFA, and *Bifidobacterium* species and SCFA, a Spearman correlation analysis was performed with a 0.05 significant level. Based on this analysis, correlation plots were created, and they are illustrated in figure 3.4.
Figure 3.4a shows that acetate is negatively correlated with almost every other SCFA. *Bacteroides* shows a strong positive correlation with isobutyric acid, whilst *Clostridium sensu stricto 1* has a strong positive correlation with the same acid, which makes these two bacteria negatively correlated with each other. *Bifidobacterium* is on this level not correlated with any of the other bacteria species or SCFAs. When *Bifidobacterium* is divided into species, shown in figure 3.4b, several correlations are present. *B. breve* shows a small negative correlation

with acetate. *B. bifidum* and *B. longum* subsp. *longum* shows a small negative correlation with isobutyric acid, whilst *B. longum* subsp. *infantis* shows a small positive correlation with the same acid. *B. longum* shows a positive correlation with acetate and is negatively correlated with propionate.

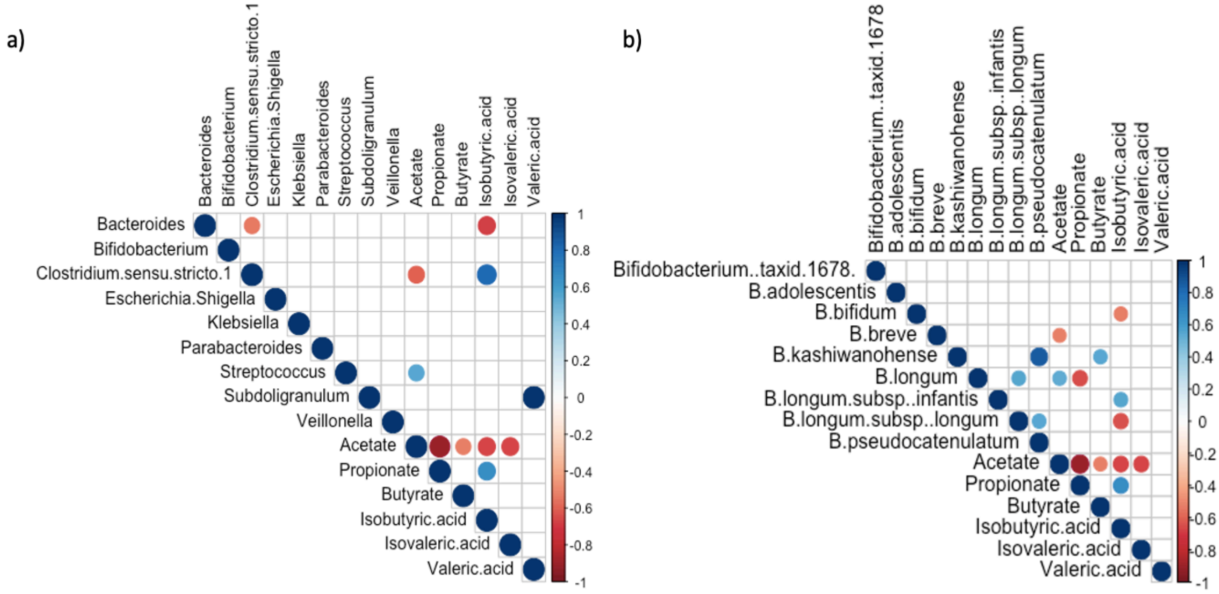
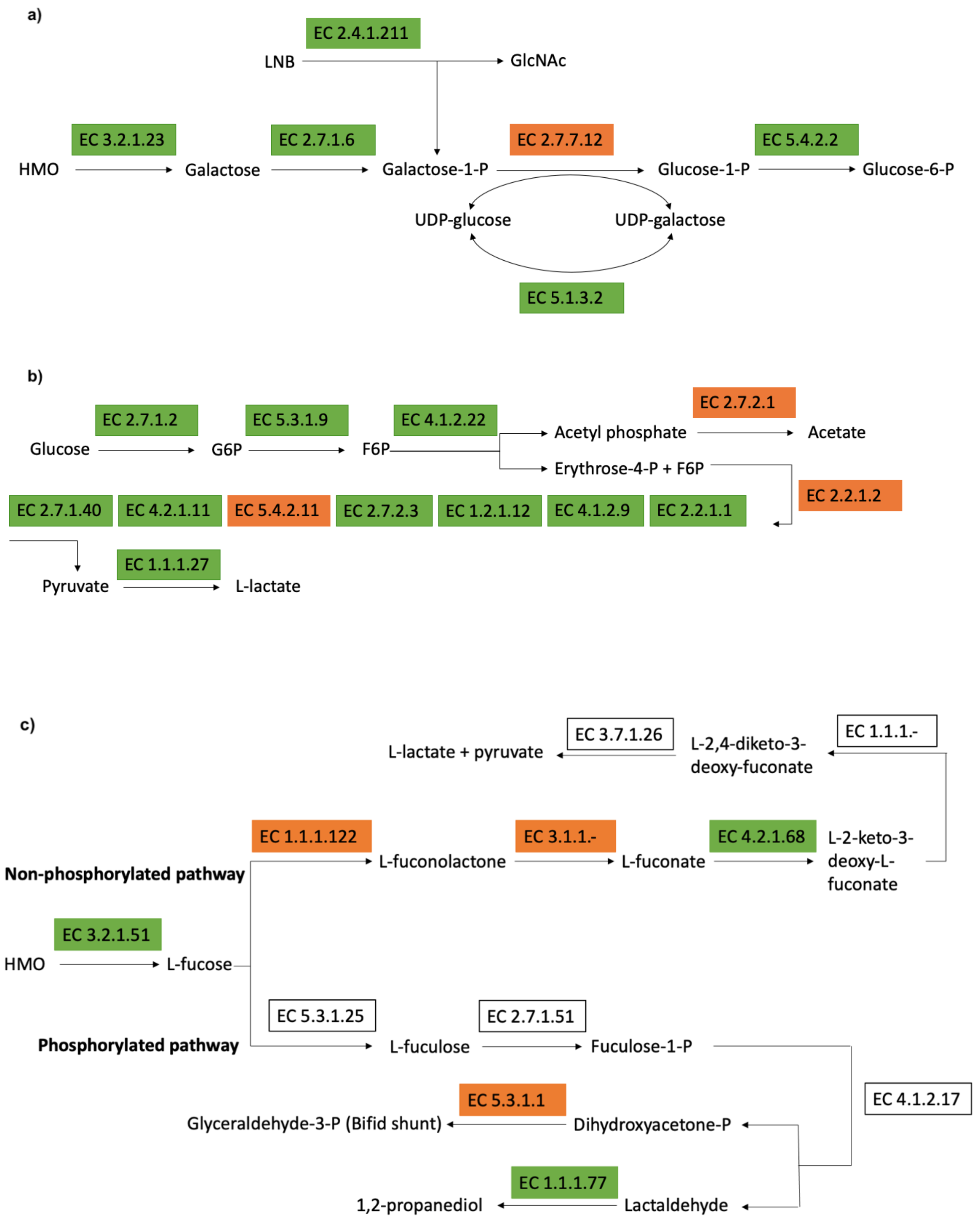


Figure 3.4: Correlation between gut bacteria and SCFAs. The figure shows significant ($p < 0.05$) correlations of the Spearman's rank correlation coefficient, rho (ρ). Figure a) shows the significant correlations between different bacterial taxa in the gut microbiota and SCFAs, whilst figure b) shows the significant correlations between *Bifidobacterium* species and SCFAs. The larger the circle, the greater the correlation, either negative or positive. Blue colors indicate degrees of positive correlation, and red colors indicate degrees of negative correlation.

3.4 HMO utilization pathways

HMO is hydrolyzed by intracellular glycoside hydrolases, which are sialidase, fucosidase, β-galactosidase and β-hexosaminidase, shown in the first boxes (after HMO) in figure 3.5a, c, d and e. All of these are found in the shotgun data, and just β-hexosaminidase (EC 3.2.1.52) is not found in the protein analysis. The two most complete pathways are the utilization of galactose and the bifid shunt (figure 3.5a and b). In these pathways, all enzymes were found in both the protein data and shotgun data, or only in the shotgun data, which means the bacteria has the ability to express the gene coding for the specific protein. The protein acetate kinase (EC 2.7.2.1), that makes acetate production possible is only found in the shotgun data, together with N-acetylglucosamine-6-phosphate deacetylase (EC 3.5.1.25) that has acetate as a biproduct. The GNB/LNB pathway is partly represented in figure 3.5a where a majority of enzymes are present in both the protein and shotgun data, and partly in figure 3.5d. The degradation pathway for GlcNAc is missing an important enzyme early in the pathway, but for extracellular GlcNAc, the pathway is complete based on the shotgun data (figure 3.5d). When degrading L-fucose, bacteria use either a pathway with or without phosphorylated intermediates, and the pathway without is more complete from the beginning, shown in figure 3.5c.



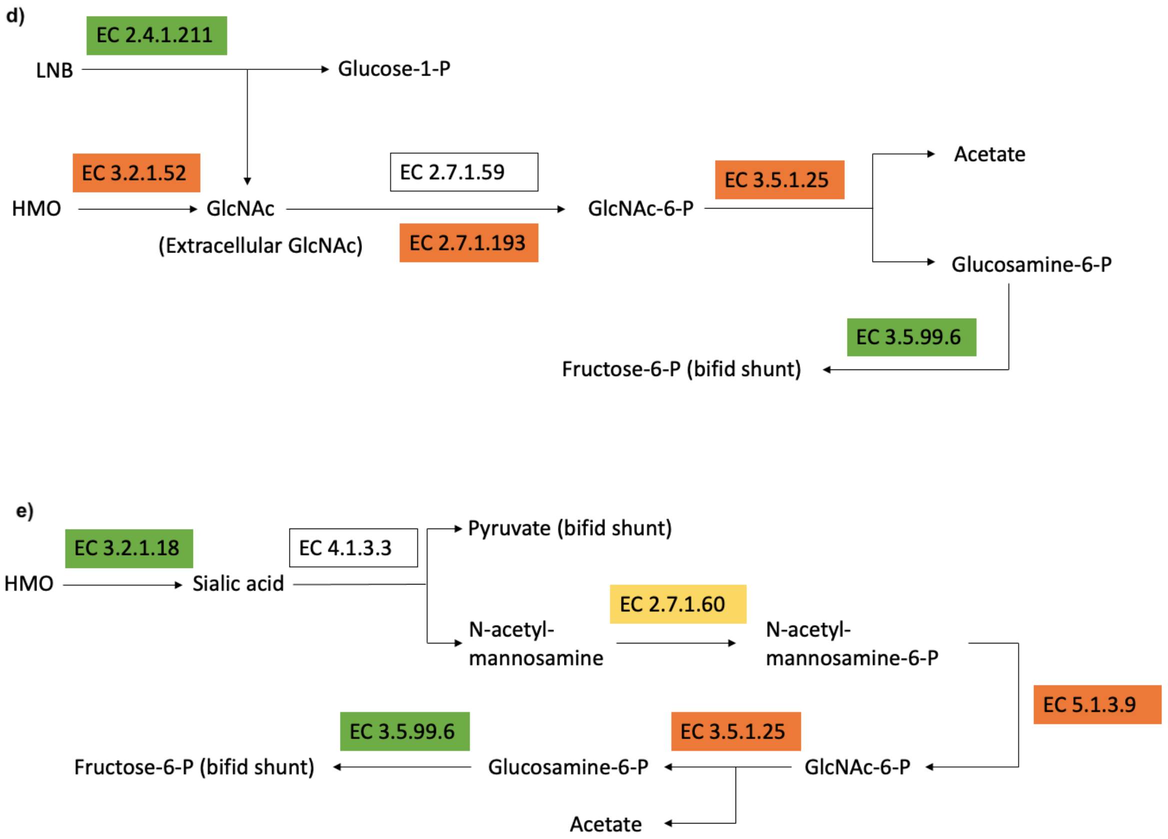


Figure 3.5: Potential HMO utilizing pathways from both proteome- and shotgun data. The proteins are shown in colored boxes with the EC number. Green boxes are proteins present in one or more samples both in the proteomics data, and in shotgun data. Orange boxes are only found in the shotgun data, and yellow boxes are only found in the proteomics data. Boxes without color are found in neither of the data analysis. The hits from the shotgun analysis are based on data from both high *Bifidobacterium* dataset 1 and 2, whilst the protein analysis is only based on data from the high *Bifidobacterium* dataset 1. The distribution of each protein in the samples, and the protein names to the EC number, are shown in more detail in the supplement, table S.5. Figure a) shows the utilization of galactose, b) shows the bifid shunt, which is a fermentative pathway found in most *Bifidobacterium* species. Figure c) shows the utilization of L-fucose, d) shows the utilization of N-acetylglucosamine and e) shows the sialic acid utilization. These pathways are shown in a more complex version in figure 1.2 and 1.3. HMO, human milk oligosaccharide; LNB, lacto-N-biose; P, phosphate; G6P, glucose-6-phosphate; F6P, fructose-6-phosphate; GlcNAc, N-acetylglucosamine.

3.5 Detection of genes central to HMO utilization

The genes expressing β-1,4-Galactosidase and 2,3/6-α-Sialidase are detected to a large extent in almost every sample (figure 3.6). LNT β-1,3-Galactosidase is also highly detected throughout the samples, but at a lower degree. The fucosidase genes are detected in the same samples, as well as the genes expressing β-N-acetylglucosaminidase.

Sample T3 consisted of 100 % *B. longum* subsp. *infantis* (figure 3.2) and was also the sample with the highest detected gene expression. Sample T4, T5 and T8 also have a lot of detected gene expression. T5 is mostly dominated by *B. longum* subsp. *infantis*, whilst T4 and T8 has most *B. longum* and a good amount *B. longum* subsp. *infantis*. On the other hand, sample T9 and T10 has high amounts of *B. longum* subsp. *infantis*, but less detected gene expression. Sample T12 does not have detection of any of the genes, and are dominated by *B. longum* and *B. bifidum*, and is also the only sample with *B. dentium*. Sample T2 is dominated by *B. breve* and has little or no detected gene expression.

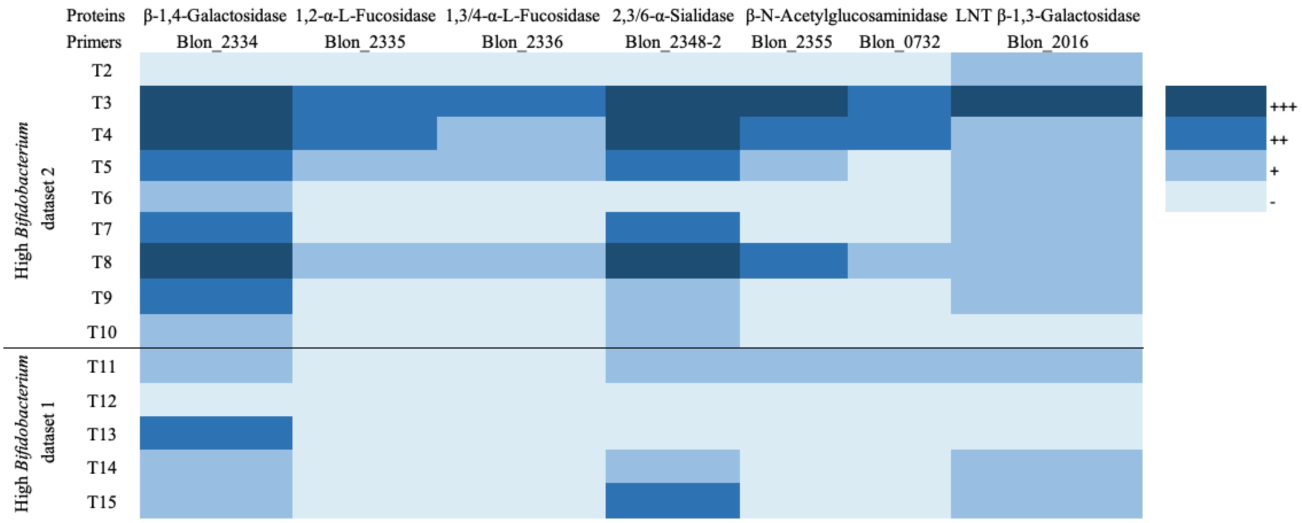


Figure 3.6: Detected gene expression linked with HMO utilizing proteins. The heatmap represents qPCR results showing gene expression in 14 samples distributed on the high *Bifidobacterium* dataset 1 and 2. The color-gradient shows degrees of differences in qPCR cycles between the control and samples. The control was parallel samples without primer during the cDNA synthesis, which will show amount of DNA created. +++ represents > 7 qPCR cycles different, ++ represents 5-7, + represents 2-5 and – represents < 2 qPCR cycles different. 2 qPCR cycles different was set as a threshold for gene expression. All primers were designed to *B. longum* subsp. *infantis*. The primer used for β-1,4-galactosidase was Blon_2334; 1,2-α-L-fucosidase, Blon_2335; 1,3/4-α-L-fucosidase, Blon_2336; 2,3/6-α-sialidase, Blon_2348-2; β-N-acetylglucosaminidase, Blon_2355 and Blon_0732; LNT β-1,3-galactosidase, Blon_2016.

4 Discussion

4.1 Potential HMO utilization pathways used by *Bifidobacterium*

To transport HMO and its derivates in through the cell membrane, the bacteria are dependent on transporters. A widely used transporter is the ATP-binding cassette (ABC) transporter. This was observed in the proteomics data, but only on a less detailed level, so the presence of an ABC transporter that is linked with HMO transport cannot be confirmed, but it is most likely present, due to the utilization proteins found inside the cell. HMO transporters are often anchored to the cell membrane or cell wall. This makes it more difficult to release them, and therefore some may have been lost during preparation for proteome analysis. This can be the reason why ABC transporters and PTSs are difficult to find from the protein data.

4.1.1 Pathways to degrade galactose and LNB

The galactose degradation pathway was one of the pathways with most enzymes present from both the shotgun data and protein data (figure 3.5). Only galactose-1-phosphate uridylyltransferase (GalT, EC 2.7.7.12) was not found in the protein data, but still found in the KEGG database from the shotgun data. Both GLNBP (EC 2.4.1.211) and galactokinase (galK, EC 2.7.1.6), which are central enzymes from the LNB/GNB pathway and Leloir pathway respectively, were highly present both in the shotgun data and protein data (figure 3.5, table S.5). This is in line with the theory from (De Bruyn et al., 2013) that *Bifidobacterium* can use both ways to degrade galactose. Which pathway the bacterium prefer is not possible to see with the analysis done in this study. Glucose-6-P produced from both of the pathways can go directly into the bifid shunt as the second substrate.

N-acetylglucosamine (GlcNAc) can be obtained from both LNB, from cleavage by GLNBP (EC 2.4.1.211), and directly from HMO, from cleavage by β-hexosaminidase (EC 3.2.1.52). GLNBP is highly present both from the protein and shotgun data (figure 3.5 and table S.5), whilst β-hexosaminidase is only found from the shotgun data. GLNBP is found in most infants associated *Bifidobacterium* species, whilst β-hexosaminidase is more species dependent. *B. longum* subsp. *infantis*, which performs an internal degradation, uses a different β-hexosaminidase to *B. bifidum*, which performs an external degradation. The enzymes used by these two species share only about 30 % identity (Garrido et al., 2012). The enzyme used to convert GlcNAc to GlcNAc-6-P, N-acetylhexosamine 1-kinase (EC 2.7.1.162), is not present in any of the two data analyzes. In the study done by (Garrido et al., 2012), several genes encoding enzymes present in the GlcNAc utilizing pathway (figure 3.5), including N-acetylhexosamine 1-kinase (EC 2.7.1.162), were expressed in *B. longum* subsp. *infantis*. N-acetyl-D-glucosamine phosphotransferase (EC 2.7.1.193), a component from the phosphotransferase system (PTS), has the ability to convert extracellular GlcNAc to GlcNAc-6-P. This enzyme exists in the shotgun data, which means that based on the shotgun data there is a complete possible utilization pathway of GlcNAc. Since GlcNAc has to be external for this pathway to be complete, *B. bifidum* may play an important role here, due to its external degradation. *B. bifidum* or other species can then make use of GlcNAc through the pathway suggested.

4.1.2 Degradation of fucose and sialic acid

L-fucose is one of the building blocks of HMO and is parted from oligosaccharides by the action of fucosidase (EC 3.2.1.51). This enzyme is present both in the shotgun data and proteome data (figure 3.5), but the next three enzymes in the phosphorylated pathway are completely missing. This is expected and similar to the results from (Bunesova et al., 2016). In the non-phosphorylated pathway, more enzymes are present from the shotgun data, compared to the phosphorylated pathway, indicating that this is the pathway used by *Bifidobacterium*. A few studies have discovered that only *B. longum* subsp. *infantis*, of the infant gut bifidobacteria, can utilize L-fucose, whilst *B. bifidum* mostly release L-fucose from HMOs (Bunesova et al., 2016; Garrido et al., 2015). The released L-fucose can then be utilized by other (bifido)bacteria by cross-feeding. In the high *Bifidobacterium* dataset 1, which was used to proteome analysis, there were few samples containing *B. longum* subsp. *infantis*, and one sample dominated by *B. bifidum*. This can explain the lack of proteins related to L-fucose degradation from the proteome analysis.
The enzyme lactaldehyde reductase (EC 1.1.1.77), catalyzing the reaction from lactaldehyde to 1,2-propanediol (1,2-PD), is present in data from both data analyzes. 1,2-PD is a precursor to the SCFA propionic acid (Bunesova et al., 2016). The presence of this enzyme indicates that some *Bifidobacterium* species can produce 1,2-PD, but the pathway in which this occurs is unknown.

Sialic acid is also left to degradation by cross-feeding by *B. bifidum*, in the same way as L-fucose (Garrido et al., 2015). According to (Egan et al., 2014), *B. breve* can cross-feed on sialic acid derivates, released by *B. bifidum*, amongst others. *B. longum* subsp. *infantis* is also thought to have sialic acid degrading abilities (Sela et al., 2008). *B. breve* is almost nonexistent in the high *Bifidobacterium* dataset 1, and the amount of *B. longum* subsp. *infantis* is minimum (figure 3.2), which can explain the lack of sialic acid degrading enzymes from the proteome analysis.

4.1.3 The bifid shunt pathway

The pathway degrading glucose, which includes the bifid shunt, has none missing enzymes. Most of the enzymes are also both found in the data from shotgun analysis and proteome analysis. According to (Turroni et al., 2018), all genes belonging to the bifid shunt are in the bifidobacterial core genome, which here means genes that are shared by all strains of *Bifidobacterium*. This theory is consistent with the result, in that the bifid shunt is the most complete pathway found based on the proteome and shotgun analyzes (figure 3.5). From the high *Bifidobacterium* dataset 1, almost every sample contained the enzymes present from the protein analysis (table S.5). The species composition in these samples are very different, which means that the bifid shunt pathway is present in several *Bifidobacterium* species associated to the infant gut microbiota.

According to (Bunesova et al., 2018; Turroni et al., 2018), the bifid shunt is a more energy yielding pathway, compared to other carbohydrate fermentative pathways found in most gut bacteria. This can contribute to the bacteria's survival capabilities, and help the genus spread faster and outcompete other carbohydrate fermentative bacteria in the gut microbiota.

4.2 Correlation between SCFA production and *Bifidobacterium*

The correlation plot between gut bacteria and SCFAs in figure 3.4a does not show any significant correlation between *Bifidobacterium* and SCFAs, but looking at species level of *Bifidobacterium*, some significant correlations with SCFAs are present (figure 3.4b). Most of the correlations are just slightly significant, but one somewhat larger is a negative correlation between *B. longum* and propionate. According to literature, *Bifidobacterium* has not been associated with propionate production, but with production of 1,2-propanediol (Bunesova et al., 2016). 1,2-propanediol is, as previously explained, a precursor to propionate and the enzyme producing 1,2-PD was found in data from both the shotgun and proteome analyzes (figure 3.5). The negative correlation is in contrast with the literature, since production of 1,2-PD should help increase, and not decrease the production of propionate in the presence of *Bifidobacterium*, as long as propionate producers are present in the gut microbiota. A possible reason for the negative correlation could be the large abundance of *Bifidobacterium* in the samples. This can cause other bacteria to be outcompeted, and some of these bacteria can be propionate producers.

4.2.1 Acetate production in *Bifidobacterium*

From the literature, is has been shown that acetate is the main SCFA produced by *Bifidobacterium* (Bunesova et al., 2018; Egan et al., 2014; LeBlanc et al., 2017). Acetate can be produced in several ways, and in connection to HMO utilization, there are two main enzymes that produce acetate: acetate kinase (EC 2.7.2.1) and N-acetylglucosamine-6-phosphate deacetylase (EC 3.5.1.25) (figure 3.5). Both of these enzymes were found in the KEGG database from the data obtained from shotgun analysis, which indicate that in the genome, the bacterium has the potential to express genes encoding acetate producing proteins. When looking at data from the protein analysis (figure 3.5), there was no discovery of either of the two acetate producing proteins. This does not exclude the potential for their existence. When performing a protein analysis, the proteins must already be expressed in the cell at the time the analysis is performed, in order to be identified. There may therefore be the case that several potential proteins found in the shotgun database, also is found in the bacteria, but are not expressed at the time the protein analysis was done.

4.3 Technical considerations and future research

4.3.1 Limitations with the proteome analysis

The preparation before the proteome analysis was done with two different methods, on one parallel of the five samples each. In one of the methods, large materials and intact human cells were removed, by passing the samples through a 20 µm filter. In the other method, filtering was done by centrifuging the samples at 1500 g for 5 min. This would collect large particles in a pellet, and the supernatant was used further in the analysis. A strength of using two different methods is that if the results are similar, the likelihood is higher that they are trustworthy, but unfortunately, one of the methods gave much lower yield. The first method performed better, giving a sufficient amount of proteins in each sample to work with, compared to the second method where only one sample could be used. The proteome analysis should therefore be performed again, using only the first method.

Another weakness with this analysis was that the sample size was too small. More samples should have been analyzed. The selection of high *Bifidobacterium* dataset 2 was done before the discovery of a lack of feces basic samples, which are used in the proteome analysis, and therefore only the high *Bifidobacterium* dataset 1 could be used for the analysis.

To get a wider understanding of the HMO degradation pathways, a proteome analysis with more samples, and samples with more variation in the *Bifidobacterium* composition should be done. This would make it easier to see differences between species, due to protein expression and HMO utilization pathways. It would be interesting to see if samples dominated by *B. breve* indeed did not have complete pathways, because they are supposedly only able to degrade HMO derivates. Samples fully dominated by *B. longum* subsp. *infantis* and *B. bifidum* would also give an indication of their utilization pattern, due to external and internal HMO degradation.

4.3.2 Limitations with the qPCR analysis for detecting gene expression

Due to low amounts of RNA, qPCR was the best option to detect any potential gene expression. The samples have been stored over several years and thawed and frozen several times. This can have made the RNA degrade, and therefore affected the amount of RNA in the samples. Another reason could be the protocol used for RNA extraction. During the preparation for qPCR analysis, the samples did not give any results on both qubit and gel electrophoresis.

All primers used in the analysis were from *B. longum* subsp. *infantis*, which most likely have affected the results. This is because the detection of gene expression in figure 3.6 was mainly found in samples with high abundance of *B. longum* subsp. *infantis*. The detection can therefore be a result of primer binding, and not actual expression, which makes these results not trustworthy.

The gene expression analysis should be reanalyzed, using a different RNA extraction protocol yielding more RNA, and then preforming RNA sequencing. Another solution could be to construct primers based on the genome data, which would be more representative for the *Bifidobacterium* genus.

5 Conclusion

Several proteins related to HMO degradation were found either expressed or with the potential to be expressed in *Bifidobacterium*. Whole degradation pathways were found for three of the building blocks for HMO; glucose, galactose and N-acetylglucosamine. In addition to this, all the main enzymes to break down HMO; β-galactosidase, fucosidase, sialidase, GLNBP and β-hexosaminidase, were found. These enzymes were all found from both the protein data and the shotgun data, except for β-hexosaminidase which was only present in the latter. This is a good indication that *Bifidobacterium* has the ability to utilize HMO and its derivates.

References

Aebersold, R. & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422 (6928): 198-207. doi: 10.1038/nature01511.

Arboleya, S., Watkins, C., Stanton, C. & Ross, R. P. (2016). Gut Bifidobacteria Populations in Human Health and Aging. *Frontiers in Microbiology*, 7 (1204). doi: 10.3389/fmicb.2016.01204.

Belizário, J. E. & Faintuch, J. (2018). Microbiome and Gut Dysbiosis. In Silvestre, R. & Torrado, E. (eds) *Metabolic Interaction in Infection*, pp. 459-476. Cham: Springer International Publishing.

Bode, L. (2006). Recent Advances on Structure, Metabolism, and Function of Human Milk Oligosaccharides. *The Journal of Nutrition*, 136 (8): 2127-2130. doi: 10.1093/jn/136.8.2127.

Bode, L. (2012). Human milk oligosaccharides: Every baby needs a sugar mama. *Glycobiology*, 22 (9): 1147-1162. doi: 10.1093/glycob/cws074.

Bunesova, V., Lacroix, C. & Schwab, C. (2016). Fucosyllactose and L-fucose utilization of infant Bifidobacterium longum and Bifidobacterium kashiwanohense. *BMC Microbiology*, 16 (1): 248. doi: 10.1186/s12866-016-0867-4.

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

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. *Scientific Reports*, 6 (1): 23129. doi: 10.1038/srep23129.

Costa, C., Giménez-Capitán, A., Karachaliou, N. & Rosell, R. (2013). Comprehensive molecular screening: from the RT-PCR to the RNA-seq. *Translational lung cancer research*, 2 (2): 87-91. doi: 10.3978/j.issn.2218-6751.2013.02.05.

Davani-Davari, D., Negahdaripour, M., Karimzadeh, I., Seifan, M., Mohkam, M., Masoumi, S. J., Berenjian, A. & Ghasemi, Y. (2019). Prebiotics: Definition, Types, Sources, Mechanisms, and Clinical Applications. *Foods (Basel, Switzerland)*, 8 (3): 92. doi: 10.3390/foods8030092.

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.

Di Gioia, D., Aloisio, I., Mazzola, G. & Biavati, B. (2014). Bifidobacteria: their impact on gut microbiota composition and their applications as probiotics in infants. *Applied Microbiology and Biotechnology*, 98 (2): 563-577. doi: 10.1007/s00253-013-5405-9.

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N. & Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences*, 107 (26): 11971. doi: 10.1073/pnas.1002601107.

Egan, M., O'Connell Motherway, M., Ventura, M. & van Sinderen, D. (2014). Metabolism of sialic acid by Bifidobacterium breve UCC2003. *Applied and environmental microbiology*, 80 (14): 4414-4426. doi: 10.1128/AEM.01114-14.

Gao, X., Jia, R., Xie, L., Kuang, L., Feng, L. & Wan, C. (2015). Obesity in school-aged children and its correlation with Gut E.coli and Bifidobacteria: a case–control study. *BMC Pediatrics*, 15 (1): 64. doi: 10.1186/s12887-015-0384-x.

Garrido, D., Ruiz-Moyano, S. & Mills, D. A. (2012). Release and utilization of N-acetyl-d-glucosamine from human milk oligosaccharides by Bifidobacterium longum subsp. infantis. *Anaerobe*, 18 (4): 430-435. doi: https://doi.org/10.1016/j.anaerobe.2012.04.012.

Garrido, D., Ruiz-Moyano, S., Lemay, D. G., Sela, D. A., German, J. B. & Mills, D. A. (2015). Comparative transcriptomics reveals key differences in the response to milk oligosaccharides of infant gut-associated bifidobacteria. *Scientific Reports*, 5 (1): 13517. doi: 10.1038/srep13517.

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

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. *Scientific Reports*, 8 (1): 13958. doi: 10.1038/s41598-018-32080-3.

Guaraldi, F. & Salvatori, G. (2012). Effect of Breast and Formula Feeding on Gut Microbiota Shaping in Newborns. *Frontiers in Cellular and Infection Microbiology*, 2 (94). doi: 10.3389/fcimb.2012.00094.

Hevia, A., Milani, C., López, P., Donado, C. D., Cuervo, A., González, S., Suárez, A., Turroni, F., Gueimonde, M., Ventura, M., et al. (2016). Allergic Patients with Long-Term Asthma Display Low Levels of Bifidobacterium adolescentis. *PLOS ONE*, 11 (2): e0147809. doi: 10.1371/journal.pone.0147809.

Hollister, E. B., Brooks, J. P. & Gentry, T. J. (2015). Nucleic Acid-Based Methods of Analysis. In *Environmental Microbiology*, pp. 280-285: Elsevier Inc.

Illumina. (2020a). Chapter 2 Protocol. In *Illumina DNA Prep Reference Guide*: Illumina.

Illumina. (2020b). *Illumina DNA Prep Reference Guide*: Illumina.

Illumina Inc. (2019). *Targeted next-generation sequencing versus qPCR and Sanger sequencing*. Illumina. Available at: https://www.illumina.com/content/dam/illumina-marketing/documents/products/other/infographic-targeted-ngs-vs-sanger-qpcr.pdf.

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

Invitrogen Corporation. (2007). *Quant-iT Assays Abbreviated Protocol*: TermoFisher. Available at: http://tools.thermofisher.com/content/sfs/manuals/mp39808.pdf.

Kitaoka, M. (2012). Bifidobacterial Enzymes Involved in the Metabolism of Human Milk Oligosaccharides. *Advances in Nutrition*, 3 (3): 422S-429S. doi: 10.3945/an.111.001420.

LeBlanc, J. G., Chain, F., Martín, R., Bermúdez-Humarán, L. G., Courau, S. & Langella, P. (2017). Beneficial effects on host energy metabolism of short-chain fatty acids and vitamins produced by commensal and probiotic bacteria. *Microbial cell factories*, 16 (1): 79-79. doi: 10.1186/s12934-017-0691-z.

Lesk, A. M. (2016). Proteomics and system biology. In *Introduction to protein science - architecture, function and genomics*, pp. 387-410: Oxford university press.

Lødrup 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.

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.

Mantovani, A., Locati, M., Vecchi, A., Sozzani, S. & Allavena, P. (2001). Decoy receptors: a strategy to regulate inflammatory cytokines and chemokines. *Trends in Immunology*, 22 (6): 328-336. doi: https://doi.org/10.1016/S1471-4906(01)01941-X.

Matsuki, T., Yahagi, K., Mori, H., Matsumoto, H., Hara, T., Tajima, S., Ogawa, E., Kodama, H., Yamamoto, K., Yamada, T., et al. (2016). A key genetic factor for fucosyllactose utilization affects infant gut microbiota development. *Nature Communications*, 7 (1): 11939. doi: 10.1038/ncomms11939.

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.

Morrison, D. J. & Preston, T. (2016). Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes*, 7 (3): 189-200. doi: 10.1080/19490976.2015.1134082.

Nilsen, M., Madelen Saunders, C., Leena Angell, I., Arntzen, M. Ø., Lødrup Carlsen, K. C., Carlsen, K.-H., Haugen, G., Heldal Hagen, L., Carlsen, M. H., Hedlin, G., et al. (2020). Butyrate Levels in the Transition from an Infant- to an Adult-Like Gut Microbiota Correlate with Bacterial Networks Associated with Eubacterium Rectale and Ruminococcus Gnavus. *Genes*, 11 (11): 1245.

O'Callaghan, A. & van Sinderen, D. (2016). Bifidobacteria and Their Role as Members of the Human Gut Microbiota. *Frontiers in Microbiology*, 7 (925). doi: 10.3389/fmicb.2016.00925.

Odamaki, T., Horigome, A., Sugahara, H., Hashikura, N., Minami, J., Xiao, J.-z. & Abe, F. (2015). Comparative Genomics Revealed Genetic Diversity and Species/Strain-Level Differences in Carbohydrate Metabolism of Three Probiotic Bifidobacterial Species. *International Journal of Genomics*, 2015: 567809. doi: 10.1155/2015/567809.

Olin, A., Henckel, E., Chen, Y., Lakshmikanth, T., Pou, C., Mikes, J., Gustafsson, A., Bernhardsson, A. K., Zhang, C., Bohlin, K., et al. (2018). Stereotypic Immune System Development in Newborn Children. *Cell*, 174 (5): 1277-1292.e14. doi: 10.1016/j.cell.2018.06.045.

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

Rodriguez, J. M., Murphy, K., Stanton, C., Ross, R. P., Kober, O. I., Juge, N., Avershina, E., Rudi, K., Narbad, A., Jenmalm, M. C., et al. (2015). The composition of the gut microbiota throughout life, with an emphasis on early life. *Microb Ecol Health Dis*, 26: 26050. doi: 10.3402/mehd.v26.26050.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74 (12): 5463. doi: 10.1073/pnas.74.12.5463.

Sela, D. A., Chapman, J., Adeuya, A., Kim, J. H., Chen, F., Whitehead, T. R., Lapidus, A., Rokhsar, D. S., Lebrilla, C. B., German, J. B., et al. (2008). The genome sequence of Bifidobacterium longum subsp. infantis reveals adaptations for milk utilization within the infant microbiome. *Proc Natl Acad Sci U S A*, 105 (48): 18964-9. doi: 10.1073/pnas.0809584105.

Sela, D. A., Li, Y., Lerno, L., Wu, S., Marcobal, A. M., German, J. B., Chen, X., Lebrilla, C. B. & Mills, D. A. (2011). An infant-associated bacterial commensal utilizes breast milk sialyloligosaccharides. *J Biol Chem*, 286 (14): 11909-18. doi: 10.1074/jbc.M110.193359.

Sela, D. A., Garrido, D., Lerno, L., Wu, S., Tan, K., Eom, H. J., Joachimiak, A., Lebrilla, C. B. & Mills, D. A. (2012). Bifidobacterium longum subsp. infantis ATCC 15697 α-fucosidases are active on fucosylated human milk oligosaccharides. *Appl Environ Microbiol*, 78 (3): 795-803. doi: 10.1128/aem.06762-11.

Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26 (10): 1135-1145. doi: 10.1038/nbt1486.

Tangerman, A. & Nagengast, F. M. (1996). A Gas Chromatographic Analysis of Fecal Short-Chain Fatty Acids, Using the Direct Injection Method. *Analytical Biochemistry*, 236 (1): 1-8. doi: https://doi.org/10.1006/abio.1996.0123.

Thermo Fisher Scientific Inc. (2018). *Qubit Assays*. ThermoFisher Scientific. Available at: https://www.thermofisher.com/no/en/home/industrial/spectroscopy-elemental-isotope-analysis/molecular-spectroscopy/fluorometers/qubit/qubit-assays.html?open=dnaqa#dnaqa.

Thursby, E. & Juge, N. (2017). Introduction to the human gut microbiota. *The Biochemical journal*, 474 (11): 1823-1836. doi: 10.1042/BCJ20160510.

Turroni, F., van Sinderen, D. & Ventura, M. (2011). Genomics and ecological overview of the genus Bifidobacterium. *International Journal of Food Microbiology*, 149 (1): 37-44. doi: https://doi.org/10.1016/j.ijfoodmicro.2010.12.010.

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

Underwood, M. A., German, J. B., Lebrilla, C. B. & Mills, D. A. (2015). Bifidobacterium longum subspecies infantis: champion colonizer of the infant gut. *Pediatric Research*, 77 (1): 229-235. doi: 10.1038/pr.2014.156.

Vitha, M. F. (2016). *Chromatography: principles and instrumentation*, vol. 185: John Wiley & Sons.

Wang, Z., Gerstein, M. & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10 (1): 57-63. doi: 10.1038/nrg2484.

Woese, C. R. & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, 74 (11): 5088. doi: 10.1073/pnas.74.11.5088.

Wong, J. M., de Souza, R., Kendall, C. W., Emam, A. & Jenkins, D. J. (2006). Colonic health: fermentation and short chain fatty acids. *J Clin Gastroenterol*, 40 (3): 235-43. doi: 10.1097/00004836-200603000-00015.

Wong, M. L. & Medrano, J. F. (2005). Real-time PCR for mRNA quantitation. *BioTechniques*, 39 (1): 75-85. doi: 10.2144/05391rv01.

Yoshida, E., Sakurama, H., Kiyohara, M., Nakajima, M., Kitaoka, M., Ashida, H., Hirose, J., Katayama, T., Yamamoto, K. & Kumagai, H. (2011). Bifidobacterium longum subsp. infantis uses two different β-galactosidases for selectively degrading type-1 and type-2 human milk oligosaccharides. *Glycobiology*, 22 (3): 361-368. doi: 10.1093/glycob/cwr116.

Yu, Y., Lee, C., Kim, J. & Hwang, S. (2005). Group-specific primer and probe sets to detect methanogenic communities using quantitative real-time polymerase chain reaction. *Biotechnology and Bioengineering*, 89 (6): 670-679. doi: https://doi.org/10.1002/bit.20347.

Supplementary tables and figures

Table S.1: Qubit results used in amplification of tagmented DNA. These values were used to calculate the number of PCR cycles.

Sample	**µg/mL**
1	1.15
2	0.128
3	0.177
4	0.172
5	0.126
6	0.130
7	0.129
8	0.223
9	0.610
10	0.204
Pos. control	2.15
Neg. control	< 0.01
11	< 0.01
12	< 0.01
13	< 0.01
14	< 0.01
15	0.062
Pos. control	1.52
Neg. control	< 0.01

Table S.2: Concentrations from BCA Protein Assay. Only one sample got a value.

Sample	**µg/mL**
11	- (too low)
11_2	-
12	-
12_2	-
13	14
14	-
14_2	-
15	-
15_2	-

Table S.3: NanoDrop measurements. The duplicate of sample 12 has been removed, due to weak appearance on SDS-PAGE.

Sample	mg/mL	A205
11	0.014	0.43
11_2	0.043	1.33
12	0.021	0.64
13	0.027	0.82
14	0.019	0.60
14_2	0.018	0.55
15	0.022	0.69
15_2	0.014	0.42

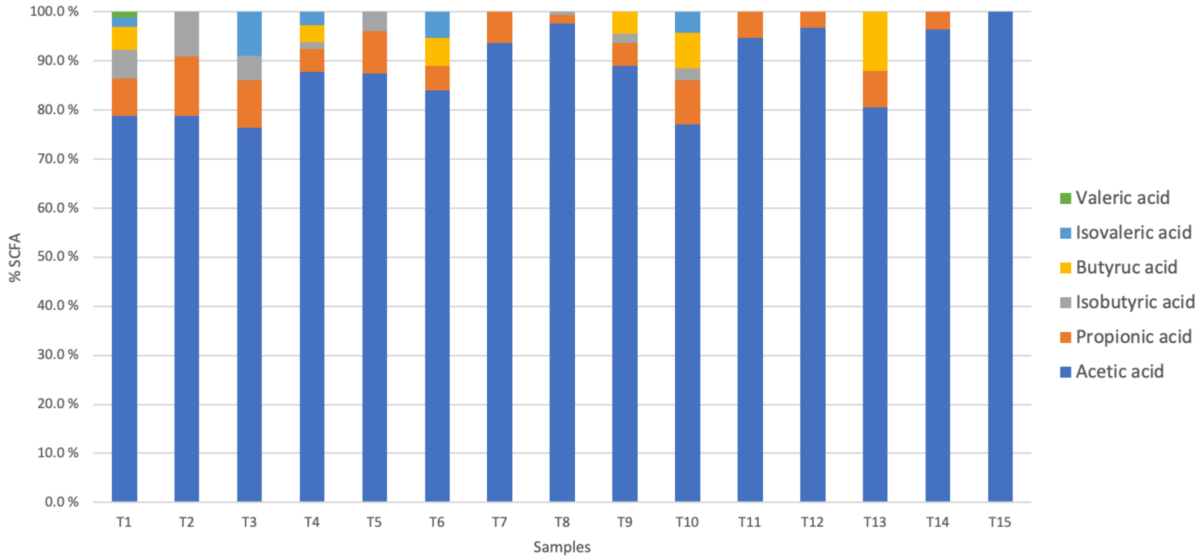


Figure S.1: Distribution of SCFAs in all 15 samples. Sample T1-T10 belongs to dataset C, whilst sample T11-T15 belongs to dataset B.

Table S.5: Presence of enzymes from proteome analysis. The table presents presence of the most known enzymes, sorted into pathways, linked with HMO utilization in the samples from proteome analysis. The data are obtained from matrices made and filtered in Perseus.

			Present in the samples				
Protein	**EC number**	**T11**	**T12**	**T13**	**T14**	**T14_2**	**T15**
L-fucose utilization							
α-L-fucosidase	EC 3.2.1.51	Yes	Yes	No	No	No	No
L-fucose isomerase	EC 5.3.1.25	No	No	No	No	No	No
L-fuculokinase	EC 2.7.1.51	No	No	No	No	No	No
L-fuculose-1-phosphate-aldolase	EC 4.1.2.17	No	No	No	No	No	No
lactaldehyde reductase	EC 1.1.1.77	Yes	Yes	Yes	Yes	Yes	Yes
triose-phosphate isomerase	EC 5.3.1.1	No	No	No	No	No	No
L-fucose dehydrogenase	EC 1.1.1.122	No	No	No	No	No	No
L-fuconolactone hydrolase	EC 3.1.1.-	No	No	No	No	No	No
L-fuconate dehydratase	EC 4.2.1.68	Yes	Yes	Yes	Yes	Yes	Yes
L-2-keto-3-deoxy-fuconate hydrolase	EC 1.1.1.-	No	No	No	No	No	No
L-2,4-diketo-3-deoxy-fuconate hydrolase	EC 3.7.1.26	No	No	No	No	No	No
Galactose utilization / Leloir pathway							
β-galactosidase	EC 3.2.1.23	Yes	Yes	No	No	No	No
galactokinase	EC 2.7.1.6	Yes	Yes	No	Yes	Yes	Yes
UDP-glucose 4-epimerase	EC 5.1.3.2	Yes	Yes	No	Yes	Yes	Yes
galactose-1-phosphate uridylyltransferase	EC 2.7.7.12	No	No	No	No	No	No
phosphoglucomutase	EC 5.4.2.2	Yes	Yes	No	Yes	Yes	Yes
N-acetylglucosamine utilization							
β-hexosaminidase (β-N-acetylhexosaminidase)	EC 3.2.1.52	No	No	No	No	No	No
N-acetylglucosamine kinase	EC 2.7.1.59	No	No	No	No	No	No
N-acetyl-D-glucosamine phosphotransferase	EC 2.7.1.193	No	No	No	No	No	No
N-acetylglucosamine-6-phosphate deacetylase	EC 3.5.1.25	No	No	No	No	No	No
glucosamine-6-phosphate isomerase	EC 3.5.99.6	Yes	Yes	Yes	Yes	Yes	Yes
Sialic acid utilization							
exo-α-sialidase	EC 3.2.1.18	Yes	Yes	No	No	No	Yes
N-acetylneuraminate lyase	EC 4.1.3.3	No	No	No	No	No	No
N-acetylmannosamine kinase	EC 2.7.1.60	Yes	Yes	No	No	No	No
N-acetylmannosamine-6-phosphate 2-epimerase	EC 5.1.3.9	No	No	No	No	No	No
GNB/LNB pathway							

GNB/LNB phosphorylase (GLNBP)	EC 2.4.1.211	Yes	Yes	Yes	Yes	Yes	Yes
N-acetylhexosamine-1-kinase	EC 2.7.1.162	No	No	No	No	No	No
UDP-glucose-hexose-1-phosphate uridylyl transferase	EC 2.7.7.12	No	No	No	No	No	No
UDP glucose/GlcNAc 4-epimerase	EC 5.1.3.2	Yes	Yes	No	Yes	Yes	Yes
Bifid shunt / central fermentative pathway							
glucokinase	EC 2.7.1.2	Yes	Yes	No	Yes	No	Yes
glucose-6-phosphate isomerase	EC 5.3.1.9	Yes	Yes	Yes	Yes	Yes	Yes
fructose-6-phosphate phosphoketolase	EC 4.1.2.22	Yes	Yes	Yes	Yes	Yes	Yes
transaldolase	EC 2.2.1.2	No	No	No	No	No	No
transketolase	EC 2.2.1.1	Yes	Yes	Yes	Yes	Yes	Yes
ribose 5-phosphate isomerase	EC 5.3.1.6	Yes	Yes	No	Yes	Yes	Yes
ribulose 5-phosphate epimerase	EC 5.1.3.4	No	No	No	No	No	No
xylulose-5-phosphate phosphoketolase	EC 4.1.2.9	Yes	Yes	Yes	Yes	Yes	Yes
acetate kinase	EC 2.7.2.1	No	No	No	No	No	No
glyceraldehyde-3-phosphate dehydrogenase	EC 1.2.1.12	Yes	Yes	Yes	Yes	Yes	Yes
phosphoglycerate kinase	EC 2.7.2.3	Yes	Yes	Yes	Yes	Yes	Yes
phosphoglycerate mutase	EC 5.4.2.11	No	No	No	No	No	No
enolase	EC 4.2.1.11	Yes	Yes	Yes	Yes	Yes	Yes
pyruvate kinase	EC 2.7.1.40	Yes	Yes	Yes	Yes	Yes	Yes
lactate dehydrogenase	EC 1.1.1.27	Yes	Yes	Yes	Yes	Yes	Yes

Table S.6: Sequence of index primers used for 16S rRNA sequencing. The table shows the 16 forward primers (F) and seven reverse primers (R) used in index PCR. The primer sequences are from (Yu et al., 2005).

Primer	**Primer sequence (5´-3´)**
F1	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctagtcaaCCTACGGGRBGCASCAG
F2	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctagttccCCTACGGGRBGCASCAG
F3	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctatgtcaCCTACGGGRBGCASCAG
F4	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctccgtccCCTACGGGRBGCASCAG
F5	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctgtagagCCTACGGGRBGCASCAG
F6	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctgtccgcCCTACGGGRBGCASCAG
F7	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctgtgaaaCCTACGGGRBGCASCAG
F8	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctgtggccCCTACGGGRBGCASCAG
F9	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctgtttcgCCTACGGGRBGCASCAG
F10	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctcgtacgCCTACGGGRBGCASCAG

F11	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctgagtggCCTACGGGRBGCASCAG
F12	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctggtagcCCTACGGGRBGCASCAG
F13	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctactgatCCTACGGGRBGCASCAG
F14	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctatgagcCCTACGGGRBGCASCAG
F15	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctattcctCCTACGGGRBGCASCAG
F16	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctcaaaagCCTACGGGRBGCASCAG
R26	caagcagaagacggcatacgagatGCTCATgtgactggagttcagacgtgtgctcttccgatctGGACTACYVGGGTATCTAAT
R27	caagcagaagacggcatacgagatAGGAATgtgactggagttcagacgtgtgctcttccgatctGGACTACYVGGGTATCTAAT
R28	caagcagaagacggcatacgagatCTTTTGgtgactggagttcagacgtgtgctcttccgatctGGACTACYVGGGTATCTAAT
R29	caagcagaagacggcatacgagatTAGTTGgtgactggagttcagacgtgtgctcttccgatctGGACTACYVGGGTATCTAAT
R30	caagcagaagacggcatacgagatCCGGTGgtgactggagttcagacgtgtgctcttccgatctGGACTACYVGGGTATCTAAT
R31	caagcagaagacggcatacgagatATCGTGgtgactggagttcagacgtgtgctcttccgatctGGACTACYVGGGTATCTAAT
R32	caagcagaagacggcatacgagatTGAGTGgtgactggagttcagacgtgtgctcttccgatctGGACTACYVGGGTATCTAAT

Appendix A: Sample information for the reference dataset for microbiota composition

Categories	**n**
Birth method	
Caesarean section	16
Vaginally	84
Gender	
Boys	46
Girls	54
Feeding method at 3 months	
Breast milk	48
Breast milk + milk powder	14
Breast milk + milk powder + complementary	19
Complementary	4
Missing	15
Feeding method at 6 months	
Breast milk	69
Breastfed from 3-6 mo., stop at 6 mo.	3
Complementary	5
Missing	23

Appendix B: GC details

Gas chromatography was done on a Trace 1310 with an autosampler (ThermoFisher Scientific, USA), with the following specifications:

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

Appendix C: LC-MS/MS program

The LC-MS/MS used was a nanoLC-Orbitrap MS/MS (Dionex Ultimate 3000 UHPLC; Thermo Scientific, Bremen, Germany), connected to a Q-Exactive mass spectrometer (Thermo Scientific, Bremen, Germany). First, peptides were applied to a trap column (Acclaim PepMap100, C18, 5 µm, 100 Å, 300 µm i.d. × 5 mm) and backflushed onto a 50 cm × 70 µm analytical column (Acclaim PepMap RSLC C18, 2 µm, 100 Å, 75 µm i.d. × 50 cm, nanoViper). A 120 min gradient from 3.2 to 36% solution B (99.9 % ACN, 0.1% formic acid) separated the proteins, at a flow rate of 300 nl/min. The Q-Exactive mass spectrometer setup was (Top5 method): a full scan (300-1600 m/z) at R=70.000 was followed by (up to) 12 MS2 scans at R=17500, using an NCE setting of 28. Singly charged precursors were excluded for MS/MS, as were precursors with z>5. Dynamic exclusion was set to 20 sec.

Appendix D: Data processing in MaxQuant

Data processing in MaxQuant was done by Ph.D Morten Nilsen. MaxQuant version 1.6.7.0 was used to analyze, identify and quantify raw files from the mass spectroscopy analysis. The algorithm used with this data program was the MaxLFQ algorithm implemented for label-free quantitative (LFQ) detection of proteins.

Raw files were searched against both the sample-specific protein sequence database and the human genome (*Homo sapiens*, 73952 sequences). The sequences database was complemented with common contaminants (human keratin, trypsin and bovine serum albumin) as well as reversed sequences of all protein entries to estimate the false discovery rate (FDR). Variable modifications were oxidation of methionine's, protein N-terminal acetylation, deamination of asparagine and glutamine, and conversion of glutamine to pyro-glutamic acid, while carbamidomethylating of cysteine residues was used as a fixed modification. Two missed cleavages of trypsin were allowed.

Appendix E: R Markdown files

E.1 Shotgun results and database for proteome analysis

Finding all unique bins with *Bifidobacterium*

First, I read the file "reads.txt" to a table with 5 columns that shows all unique sequences that exists in all bins, and which taxonomy and number of bp these have. The different columns stand for:

- C/U = classified/unclassified
- Seq.ID = from which sequence the taxonomy comes from
- Tax.ID = which taxonomy that has the sequence
- bp.length = how many basepairs that makes the sequence
- LCA = refers to LCA classifier

```
library(tidyverse)
library(dplyr)

reads_file <- "reads.txt"

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

##
## ── Column specification ───────────────────────────────────────────────
─────────
## cols(
##   `C/U` = col_character(),
##   Seq.ID = col_character(),
##   Tax.ID = col_character(),
##   bp.length = col_double(),
##   LCA = col_character()
## )

head(new_krk.tbl, 3) #Shows the first 3 rows of the table to check if all
is correct

## # A tibble: 3 x 5
##   `C/U` Seq.ID           Tax.ID           bp.length LCA
##   <chr> <chr>            <chr>                <dbl> <chr>
## 1 C     NODE_6_length_2… Collinsella aer…    215984 84999:45 0:3 84999:
5 0:57 7…
## 2 C     NODE_12_length_… Collinsella aer…    174033 84999:28 2003188:6
84999:4 …
## 3 C     NODE_16_length_… Collinsella aer…    150145 0:61 34:5 0:325 34:
5 0:2718…
```

To easily access the files in each bin, I listed all files in a vector from a folder with all bins. Then I made a loop that collect all Seq.IDs from every bin in the folder and connects these to the right bin in a table “df”.

```
files <- list.files(path = "All_bins/", pattern = "*.fa", full.names = T)

df <- data.frame(matrix(ncol = 2, dimnames = list(NULL, c("Seq.ID", "Bin_i
d"))))

for (i in 1:length(files)) {
  lines1 <- readLines(files[i])
  logicals1 <- str_detect(lines1, pattern = ">")
  idx1 <- which(logicals1)
  fasta.tbl <- tibble(Seq.ID = lines1[idx1])
  fasta.tbl$Bin_id <- rep(files[i], nrow(fasta.tbl))
  df <- rbind.data.frame(df, fasta.tbl)
}

df <- df[-1,] #Have to remove the first row, because this is blank

head(df, 3)

##                                  Seq.ID                 Bin_id
## 2  >NODE_6_length_215984_cov_97.521884 All_bins//M1.001.fasta
## 3 >NODE_12_length_174033_cov_81.709298 All_bins//M1.001.fasta
## 4 >NODE_16_length_150145_cov_83.615665 All_bins//M1.001.fasta
```

In the column Seq.ID in the table “df” a “>” is before the IDs, and this does not exist in new_krk.tbl. The “>” must therefore be removed from “df” before the two tables can be merged. Merged df and new_krk.tbl to connect taxonomy to wanted samples.

```
df$Seq.ID <- df$Seq.ID %>% gsub(pattern = ">", replacement = "")

new.table <- left_join(new_krk.tbl, df, by = "Seq.ID")

head(new.table, 3)

## # A tibble: 3 x 6
##   `C/U` Seq.ID         Tax.ID         bp.length LCA
Bin_id
##   <chr> <chr>          <chr>              <dbl> <chr>
<chr>
## 1 C     NODE_6_length… Collinsella a…    215984 84999:45 0:3 84999:5 …
All_bins…
## 2 C     NODE_12_lengt… Collinsella a…    174033 84999:28 2003188:6 84…
All_bins…
## 3 C     NODE_16_lengt… Collinsella a…    150145 0:61 34:5 0:325 34:5 …
All_bins…
```

Extracted all *Bifidobacterium* species from my samples (Sample and T) in their own table and bound them together to “tbl”.

```
tbl1 <- new.table %>%
  filter(str_detect(Tax.ID, "Bifidobacterium")) %>%
  filter(str_detect(Bin_id, "Sample"))

tbl2 <- new.table  %>%
  filter(str_detect(Tax.ID, "Bifidobacterium")) %>%
  filter(str_detect(Bin_id, "T"))

only_bifido <- bind_rows(tbl1, tbl2)

head(only_bifido, 3)

## # A tibble: 3 x 6
##   `C/U` Seq.ID        Tax.ID          bp.length LCA                   B
in_id
##   <chr> <chr>         <chr>               <dbl> <chr>                 <
chr>
## 1 C     NODE_9601_le… Bifidobacteriu…       955 1678:97 1685:21 1678… A
ll_bins/…
## 2 C     NODE_10649_l… Bifidobacteriu…       875 1678:11 216816:12 16… A
ll_bins/…
## 3 C     NODE_3323_le… Bifidobacteriu…      2442 1678:232 1685:23 0:3… A
ll_bins/…
```

Extracted unique bins from the table only containing *Bifidobacterium.*

```
unique.bins <- unique(only_bifido$Bin_id)
print(unique.bins)

##  [1] "All_bins//Sample10.005.fasta" "All_bins//Sample10.006.fasta"
##  [3] "All_bins//Sample10.008.fasta" "All_bins//Sample10.013.fasta"
##  [5] "All_bins//Sample1.003.fasta"  "All_bins//Sample1.004.fasta"
##  [7] "All_bins//Sample1.005.fasta"  "All_bins//Sample1.008.fasta"
##  [9] "All_bins//Sample10.2.fa"      "All_bins//Sample1.21.fa"
## [11] "All_bins//Sample1.8.fa"       "All_bins//Sample2.001.fasta"
## [13] "All_bins//Sample2.005.fasta"  "All_bins//Sample2.008.fasta"
## [15] "All_bins//Sample2.009.fasta"  "All_bins//Sample2.012.fasta"
## [17] "All_bins//Sample2.013.fasta"  "All_bins//Sample3.003.fasta"
## [19] "All_bins//Sample4.003.fasta"  "All_bins//Sample4.005.fasta"
## [21] "All_bins//Sample5.001.fasta"  "All_bins//Sample5.002.fasta"
## [23] "All_bins//Sample6.005.fasta"  "All_bins//Sample6.006.fasta"
## [25] "All_bins//Sample6.007.fasta"  "All_bins//Sample6.012.fasta"
## [27] "All_bins//Sample7.001.fasta"  "All_bins//Sample7.003.fasta"
## [29] "All_bins//Sample7.004.fasta"  "All_bins//Sample7.005.fasta"
## [31] "All_bins//Sample7.006.fasta"  "All_bins//Sample7.009.fasta"
## [33] "All_bins//Sample8.006.fasta"  "All_bins//Sample9.001.fasta"
## [35] "All_bins//Sample9.002.fasta"  "All_bins//T10.003.fasta"
## [37] "All_bins//T10.004.fasta"      "All_bins//T6.003.fasta"
## [39] "All_bins//T6.004.fasta"       "All_bins//T6.006.fasta"
## [41] "All_bins//T7.005.fasta"       "All_bins//T7.007.fasta"
## [43] "All_bins//T7.009.fasta"       "All_bins//T7.010.fasta"
## [45] "All_bins//T7.016.fasta"       "All_bins//T8.008.fasta"
## [47] "All_bins//T8.011.fasta"       "All_bins//T8.012.fasta"
## [49] "All_bins//T8.24.fa"           "All_bins//T9.005.fasta"
## [51] "All_bins//T9.007.fasta"
```

Making fasta file with the sequence ID from only *Bifidobacterium* species and their predicted amino acid sequences

With the unique bins containing DNA sequence, I got corresponding bins with estimated amino acid sequence.

To read in the amino acid sequences to a table, I made a loop that extract all sequences from the folder “aasequence”, both Seq.ID and the belonging sequence, and attached to right bin. This was done by combining two loops. The first loop was to read all files in the folder, as done before, and the second was to get all sequences in their own column corresponding to their Seq.ID. First, I made a vector with the path to get the sequences from the folder.

```
library(tidyverse)
library(dplyr)

bins <- list.files(path = "aasequence/", pattern = "*.fa", full.names = T)

df_table <- data.frame(matrix(ncol = 3,
                              dimnames = list(NULL, c("Seq.ID", "Sequence"
, "Bin_id"))))

for (b in 1:length(bins)) {
  lines3 <- readLines(bins[b])
  idx3 <- which(str_detect(lines3, pattern = ">"))
  fasta.tbl3 <- tibble(Seq.ID = lines3[idx3], Sequence = "", Bin_id = "")
  N.rows <- nrow(fasta.tbl3)
  for (row in 1:N.rows) {
    seq.line.first <- idx3[row] + 1
    if(row == N.rows){
      seq.line.last <- length(lines3)
    } else {
      seq.line.last <- idx3[row + 1] - 1
    }
    seq.lines <- lines3[seq.line.first:seq.line.last]
    fasta.tbl3$Sequence[row] <- str_c(seq.lines, collapse = "")
  }
  fasta.tbl3$Bin_id <- rep(bins[b], nrow(fasta.tbl3))
  df_table <- rbind.data.frame(df_table, fasta.tbl3)
}

df_table <- df_table[-1,] #Have to remove the first row, because this is b
```

```
Lank

head(df_table, 3)

##
Seq.ID
## 2     >NODE_1_length_450894_cov_23.679890_1 # 2 # 529 # 1 # ID=1_1;part
ial=10;start_type=Edge;rbs_motif=None;rbs_spacer=None;gc_cont=0.472
## 3 >NODE_1_length_450894_cov_23.679890_2 # 716 # 1339 # 1 # ID=1_2;parti
al=00;start_type=ATG;rbs_motif=AGGA;rbs_spacer=5-10bp;gc_cont=0.458
## 4 >NODE_1_length_450894_cov_23.679890_3 # 1391 # 1738 # -1 # ID=1_3;par
tial=00;start_type=ATG;rbs_motif=None;rbs_spacer=None;gc_cont=0.526
##
Sequence
## 2                                GHCGYALKATHVPNSTGYFRCTKRTENKGCPGCGKIR
KEEFEQFIFSTMQEKFKDFQILHGREEKVNPKLTAYQVELAQVEAEIEKLLDTLTGANATLLAYANKKIEELDT
RRQTISKAIAELSVETISPQQIKKLSYYLDNWDSIDFDDKRKAADGLISTIKATSDRVQIEWKI*
## 3 MAKKNTKRGFTLVELIVVLVILAILAALLIPALTGYIDKARKSQVVAETRMLTQAVQTEMSTLYASNEY
ATLLKVGKNAFTAAAKGGQPVFDYERQLTSLAERYNAIVKLSEVPSLSDGSGSFFAVANYKCQLKWVVYSDGKG
YYGVYCQADGTVTGYSNKEVTGYETYYDTNIGKVICDVTADVTDPDEPVPWTKTAVYYGLGLLN*
## 4
MYDGDKLISFVDGFVTDDADLTDEMYENAAMHNENGAWQMIFGVNTLPAYRQQGYAGELIQKAITDAKEQGRKG
LVLTCKNRLVHYYARFGFVDEGMTDKSTHGNVAWHQMRLAF*
##                               Bin_id
## 2 aasequence//Sample1.003.fasta.faa
## 3 aasequence//Sample1.003.fasta.faa
## 4 aasequence//Sample1.003.fasta.faa
```

Removed “>” from the Seq.ID and “.faa” from the Bin_id to be able to combine df_table with only_bifido by both Seq.ID and Bin_id. By combining these tables I will get a table with amino acid sequences only attached to the wanted Seq.IDs with *Bifidobacterium*.

```
df_table$Seq.ID <- df_table$Seq.ID %>% gsub(pattern = ">", replacement = "
")
df_table$Bin_id <- df_table$Bin_id %>% gsub(pattern = ".faa", replacement
= "")

head(df_table, 3)

##
Seq.ID
## 2     NODE_1_length_450894_cov_23.679890_1 # 2 # 529 # 1 # ID=1_1;parti
al=10;start_type=Edge;rbs_motif=None;rbs_spacer=None;gc_cont=0.472
## 3 NODE_1_length_450894_cov_23.679890_2 # 716 # 1339 # 1 # ID=1_2;partia
l=00;start_type=ATG;rbs_motif=AGGA;rbs_spacer=5-10bp;gc_cont=0.458
## 4 NODE_1_length_450894_cov_23.679890_3 # 1391 # 1738 # -1 # ID=1_3;part
ial=00;start_type=ATG;rbs_motif=None;rbs_spacer=None;gc_cont=0.526
##
Sequence
## 2                                GHCGYALKATHVPNSTGYFRCTKRTENKGCPGCGKIR
KEEFEQFIFSTMQEKFKDFQILHGREEKVNPKLTAYQVELAQVEAEIEKLLDTLTGANATLLAYANKKIEELDT
RRQTISKAIAELSVETISPQQIKKLSYYLDNWDSIDFDDKRKAADGLISTIKATSDRVQIEWKI*
```

```
## 3 MAKKNTKRGFTLVELIVVLVILAILAALLIPALTGYIDKARKSQVVAETRMLTQAVQTEMSTLYASNEY
ATLLKVGKNAFTAAAKGGQPVFDYERQLTSLAERYNAIVKLSEVPSLSDGSGSFFAVANYKCQLKWVVYSDGKG
YYGVYCQADGTVTGYSNKEVTGYETYYDTNIGKVICDVTADVTDPDEPVPWTKTAVYYGLGLLN*
## 4
MYDGDKLISFVDGFVTDDADLTDEMYENAAMHNENGAWQMIFGVNTLPAYRQQGYAGELIQKAITDAKEQGRKG
LVLTCKNRLVHYYARFGFVDEGMTDKSTHGNVAWHQMRLAF*
##                             Bin_id
## 2 aasequence//Sample1.003.fasta
## 3 aasequence//Sample1.003.fasta
## 4 aasequence//Sample1.003.fasta
```

Made a loop to extract all sequences connected to Seq.ID from “only_bifido”. Included Tax.ID here to get an overview over the different *Bifidobacterium* species.

```
aa_bifido_seq_tbl <- data.frame(matrix(ncol = 4, dimnames = list(NULL, c("
Seq.ID", "Sequence", "Bin_id", "Tax.ID"))))

for (c in 1:nrow(only_bifido)) {
  indeks <- grep(only_bifido$Seq.ID[c], df_table$Seq.ID)
  bifido_tbl <- tibble(Seq.ID = df_table$Seq.ID[indeks],
                Sequence = df_table$Sequence[indeks],
                Tax.ID = only_bifido$Tax.ID[c])
  bifido_tbl$Bin_id <- rep(only_bifido$Bin_id[c], nrow(bifido_tbl))
  aa_bifido_seq_tbl <- rbind.data.frame(aa_bifido_seq_tbl, bifido_tbl)
}

aa_bifido_seq_tbl <- aa_bifido_seq_tbl[-1,] #Have to remove the first row,
because this is blank

head(aa_bifido_seq_tbl, 3)

##
Seq.ID
## 2 NODE_9601_length_955_cov_23.806075_1 # 268 # 702 # -1 # ID=101_1;part
ial=00;start_type=ATG;rbs_motif=AGGAG;rbs_spacer=5-10bp;gc_cont=0.494
## 3    NODE_10649_length_875_cov_38.738402_1 # 2 # 874 # -1 # ID=103_1;pa
rtial=11;start_type=Edge;rbs_motif=None;rbs_spacer=None;gc_cont=0.684
## 4    NODE_466_length_21106_cov_18.691484_1 # 67 # 672 # 1 # ID=40_1;par
tial=00;start_type=ATG;rbs_motif=AGGA;rbs_spacer=5-10bp;gc_cont=0.543
##
Sequence
## 2
MASLLQGFKEVQLVKPVGKTVMTVTDSVVRFNKATAEVLNFPAQVKILINDKTRQIAVTPTTAKADNAVKFSKG
EGKQTTSVSIKDAVLVEAISKYFTLVEAPEGEVSFASANGTAYPEDKTVIFDVANATAGTMKRRGRKKAE*
## 3 AYIDEFKDRFRVGPICRVLAASLDCGSVTPRGYRMFRSRPVSRMAARHEALARDILEIHADSFMAVYGY
RKTRARLLARGWDPAEIGRDQVTNVMRELGIRGVRRGGTPVATEPAKGTGGRPDLVERRFEAEAPNRLHVADIT
YVRMANGSFGYTAFAADVFARRIVGWACATTLDTRELPLQALEQAISWAASHGGADGLVHHSDHGAQYISLVYT
TRVGEFGMLPSTGTVGDSYDNAMAESADGAYKTELVWRRKPFQDSRDLESATFRWVSWRGLEASAPVLGLQDTG
## 4
MHIMFVCTGNICRSPMGELLLTRYLSGTTVQVSSAGTHGLPMHQIDPNSALLMESVGIEPSGFRSRRLTQPMAK
SADLILCFEKDQRKDIVTLAPTAVKYTFLLGDFANMCEYCARNGLVKGLTIQERLQSVINSSSIIRPMLPEPED
IEDPHGKEYAKFRTAAEQTNKALRTILTSMRKHYRVEEAPVRPQITRQYAYTV*
```

```
##                         Bin_id                                        T
ax.ID
## 2 All_bins//Sample10.005.fasta            Bifidobacterium breve (taxid
1685)
## 3 All_bins//Sample10.005.fasta Bifidobacterium longum NCC2705 (taxid 20
6672)
## 4 All_bins//Sample10.008.fasta    Bifidobacterium breve 689b (taxid 138
5942)
```

Removed "*" from the amino acid sequence, because this will interfere with programs used later in the data analysis.

```
aa_bifido_seq_tbl$Sequence <- aa_bifido_seq_tbl$Sequence %>%
  gsub(pattern = "\\*", replacement = "")
```

Removed duplicates from the dataset to only get unique amino acid sequences. This was done because I got some duplicates in the Seq.ID column that prevented the finished datafile to go through the KEGG-database.

```
new_aa_bifido_seq_tbl <- subset(aa_bifido_seq_tbl, !duplicated(aa_bifido_s
eq_tbl$Sequence))
```

Created a fasta file that contains Seq.ID and amino acid sequence from only *Bifidobacterium* species from my samples, to use as a reference in the proteome analysis.

```
library(ampir)

df_to_faa(new_aa_bifido_seq_tbl, file = "Bifido_aa_contigs.fasta")

head(new_aa_bifido_seq_tbl, 3)

##
Seq.ID
## 2 NODE_9601_length_955_cov_23.806075_1 # 268 # 702 # -1 # ID=101_1;part
ial=00;start_type=ATG;rbs_motif=AGGAG;rbs_spacer=5-10bp;gc_cont=0.494
## 3    NODE_10649_length_875_cov_38.738402_1 # 2 # 874 # -1 # ID=103_1;pa
rtial=11;start_type=Edge;rbs_motif=None;rbs_spacer=None;gc_cont=0.684
## 4    NODE_466_length_21106_cov_18.691484_1 # 67 # 672 # 1 # ID=40_1;par
tial=00;start_type=ATG;rbs_motif=AGGA;rbs_spacer=5-10bp;gc_cont=0.543
##
Sequence
## 2
MASLLQGFKEVQLVKPVGKTVMTVTDSVVRFNKATAEVLNFPAQVKILINDKTRQIAVTPTTAKADNAVKFSKG
EGKQTTSVSIKDAVLVEAISKYFTLVEAPEGEVSFASANGTAYPEDKTVIFDVANATAGTMKRRGRKKAE
## 3 AYIDEFKDRFRVGPICRVLAASLDCGSVTPRGYRMFRSRPVSRMAARHEALARDILEIHADSFMAVYGY
RKTRARLLARGWDPAEIGRDQVTNVMRELGIRGVRRGGTPVATEPAKGTGGRPDLVERRFEAEAPNRLHVADIT
YVRMANGSFGYTAFAADVFARRIVGWACATTLDTRELPLQALEQAISWAASHGGADGLVHHSDHGAQYISLVYT
TRVGEFGMLPSTGTVGDSYDNAMAESADGAYKTELVWRRKPFQDSRDLESATFRWVSWRGLEASAPVLGLQDTG
```

```
## 4
MHIMFVCTGNICRSPMGELLLTRYLSGTTVQVSSAGTHGLPMHQIDPNSALLMESVGIEPSGFRSRRLTQPMAK
SADLILCFEKDQRKDIVTLAPTAVKYTFLLGDFANMCEYCARNGLVKGLTIQERLQSVINSSSIIRPMLPEPED
IEDPHGKEYAKFRTAAEQTNKALRTILTSMRKHYRVEEAPVRPQITRQYAYTV
##                         Bin_id                                        T
ax.ID
## 2 All_bins//Sample10.005.fasta            Bifidobacterium breve (taxid
1685)
## 3 All_bins//Sample10.005.fasta Bifidobacterium longum NCC2705 (taxid 20
6672)
## 4 All_bins//Sample10.008.fasta    Bifidobacterium breve 689b (taxid 138
5942)
```

E.2 Correlation analysis

Statistical analysis with correlation plots

First, I loaded in the 16S, shotgun and SCFAs data, before I put 16S and SCFAs together in one table and shotgun and SCFAs together in another table. Then I did a correlation analysis within each table and based on this analysis, I made correlation plots.

```
library(tidyverse)
library(dplyr)

results_16S <- read.delim("data/ res_16S_utvalg.txt")

results_SCFA <- read.delim("data/res_SCFA.txt")

result_Bifido <- read.delim("data/res_Bifido.txt")

#Makes the samples in the first column of each table be the incorporated f
irst column. This will make the tables only consisting of values.
row.names(results_16S) <- results_16S$Sample
results_16S[1] <- NULL

row.names(result_Bifido) <- result_Bifido$Samples
result_Bifido[1] <- NULL

row.names(results_SCFA) <- results_SCFA$Sample
results_SCFA[1] <- NULL


#Removed two species (columns) from the table with shotgun data because th
ey had only one value, and this was very low compared to the values of the
other species.
results_Bifido_reduced <- result_Bifido %>%
  select(Bifidobacterium..taxid.1678.,
         B.adolescentis,
         B.bifidum,
         B.breve,
         B.kashiwanohense,
         B.longum,
         B.longum.subsp..infantis,
         B.longum.subsp..longum,
         B.pseudocatenulatum)

#Combined the factors to be correlated into the same table, and printing t
he first 3 rows to see that everything is correct
df_Bifido.red_SCFA <- bind_cols(results_Bifido_reduced, results_SCFA)
head(df_Bifido.red_SCFA, 3)
```

```
##    Bifidobacterium..taxid.1678. B.adolescentis  B.bifidum   B.breve
## T1                   0.00000000    0.000000000 0.01754386 0.9122807
## T2                   0.02016129    0.004032258 0.03225807 0.7661290
## T3                   0.00000000    0.000000000 0.00000000 0.0000000
##    B.kashiwanohense   B.longum B.longum.subsp..infantis B.longum.subsp.
.longum
## T1                0 0.00000000               0.05263158             0.0
1754386
## T2                0 0.02016129               0.14516129             0.0
1209677
## T3                0 0.00000000               1.00000000             0.0
0000000
##    B.pseudocatenulatum Acetate Propionate Butyrate Isobutyric.acid
## T1                   0 0.78816    0.07561  0.04585         0.06001
## T2                   0 0.78857    0.12144  0.00000         0.08999
## T3                   0 0.76474    0.09666  0.00000         0.04879
##    Isovaleric.acid Valeric.acid
## T1         0.01868      0.01169
## T2         0.00000      0.00000
## T3         0.08980      0.00000
```

```
df_16S_SCFA <- bind_cols(results_16S, results_SCFA)
head(df_16S_SCFA, 3)
```

```
##    Bacteroides Bifidobacterium Clostridium.sensu.stricto.1 Escherichia.
Shigella
## T1           0         0.39282                     0.06825
0.09076
## T2           0         0.60856                     0.10716
0.07769
## T3           0         0.79012                     0.01023
0.14831
##    Klebsiella Parabacteroides Streptococcus Subdoligranulum Veillonella
Acetate
## T1    0.00000               0             0         0.20775     0.04804
0.78816
## T2    0.10209               0             0         0.00000     0.03267
0.78857
## T3    0.00000               0             0         0.00000     0.00000
0.76474
##    Propionate Butyrate Isobutyric.acid Isovaleric.acid Valeric.acid
## T1    0.07561  0.04585         0.06001         0.01868      0.01169
## T2    0.12144  0.00000         0.08999         0.00000      0.00000
## T3    0.09666  0.00000         0.04879         0.08980      0.00000
```

The correlation analysis used was Spearman correlation. This resulted in a list of 4, “R”, “P”, “P.unadj” and “type”, where “R” consist of a list of 3. Inside “R” we find a “r” matrix which shows Spearman correlation, a “n” matrix which shows the number of observations and a “P” matrix which shows the p-values (pairwise two-sided p-values).

```
library(RcmdrMisc)

corr_16S_SCFA <- rcorr.adjust(df_16S_SCFA, type = c("spearman"))

corr_Bifido.red_SCFA <- rcorr.adjust(df_Bifido.red_SCFA, type = c("spearma
n"))
```

In the correlation plots, both the Spearman correlation matrix and the pairwise two-sided p-values matrix was used.

```
head(corr_16S_SCFA$R$r, 3)

##                             Bacteroides Bifidobacterium
## Bacteroides                   1.0000000      -0.0786700
## Bifidobacterium              -0.0786700       1.0000000
## Clostridium.sensu.stricto.1  -0.5397065      -0.3021836
##                             Clostridium.sensu.stricto.1 Escherichia.Shi
gella
## Bacteroides                                  -0.5397065          -0.328
25514
## Bifidobacterium                              -0.3021836           0.324
33749
## Clostridium.sensu.stricto.1                   1.0000000           0.083
34435
##                             Klebsiella Parabacteroides Streptococcus
## Bacteroides                  0.2264853       0.2946172   0.062584482
## Bifidobacterium             -0.3300115      -0.2474358   0.007377111
## Clostridium.sensu.stricto.1  0.0000000      -0.2093589  -0.474382716
##                             Subdoligranulum Veillonella   Acetate Propi
onate
## Bacteroides                      -0.1841357  -0.3432734  0.206243 -0.09
14273
## Bifidobacterium                  -0.4330127  -0.2396494  0.300000 -0.28
57143
## Clostridium.sensu.stricto.1       0.4187179   0.3354102 -0.596309  0.47
94647
##                                Butyrate Isobutyric.acid Isovaleric.acid
## Bacteroides                  0.06956217      -0.6723686      -0.1949367
## Bifidobacterium             -0.31427096      -0.1430554      -0.2636508
## Clostridium.sensu.stricto.1  0.29545455       0.8069414       0.3598043
##                             Valeric.acid
## Bacteroides                   -0.1841357
## Bifidobacterium               -0.4330127
## Clostridium.sensu.stricto.1    0.4187179
```

```
head(corr_16S_SCFA$R$P, 3)
```

```
##                               Bacteroides Bifidobacterium
## Bacteroides                            NA       0.7804857
## Bifidobacterium                0.78048570              NA
## Clostridium.sensu.stricto.1    0.03784261       0.2736630
##                               Clostridium.sensu.stricto.1 Escherichia.Shi
gella
## Bacteroides                                    0.03784261            0.23
22862
## Bifidobacterium                                0.27366296            0.23
82386
## Clostridium.sensu.stricto.1                            NA            0.76
77636
##                               Klebsiella Parabacteroides Streptococcus
## Bacteroides                    0.4169609       0.2864478    0.82464366
## Bifidobacterium                0.2296480       0.3739394    0.97918334
## Clostridium.sensu.stricto.1    1.0000000       0.4539422    0.07399861
##                               Subdoligranulum Veillonella    Acetate Prop
ionate
## Bacteroides                         0.5112198   0.2103339 0.46084019 0.74
589527
## Bifidobacterium                     0.1069075   0.3896377 0.27731678 0.30
193635
## Clostridium.sensu.stricto.1         0.1203291   0.2216564 0.01896201 0.07
052975
##                                Butyrate Isobutyric.acid Isovaleric.acid
## Bacteroides                   0.8054207    0.0060301732       0.4862953
## Bifidobacterium               0.2539635    0.6110227517       0.3423781
## Clostridium.sensu.stricto.1 0.2850158    0.0002769747       0.1877498
##                               Valeric.acid
## Bacteroides                      0.5112198
## Bifidobacterium                  0.1069075
## Clostridium.sensu.stricto.1    0.1203291
```

```
head(corr_Bifido.red_SCFA$R$r, 3)
```

```
##                                Bifidobacterium..taxid.1678. B.adolescenti
s
## Bifidobacterium..taxid.1678.                      1.0000000     0.2781296
3
## B.adolescentis                                    0.2781296     1.0000000
0
## B.bifidum                                         0.1789697     0.0250477
5
##                                 B.bifidum    B.breve B.kashiwanohense
B.longum
## Bifidobacterium..taxid.1678. 0.17896966 0.36159386       0.18555425  0.
09815714
## B.adolescentis                0.02504775 0.11438801       0.03931574 -0.
03074458
## B.bifidum                     1.00000000 0.05822174       0.16805028 -0.
24802060
##                                B.longum.subsp..infantis B.longum.subsp..l
ongum
```

```
## Bifidobacterium..taxid.1678.               0.08108633            0.0490
78569
## B.adolescentis                             0.02305844            0.0076
86145
## B.bifidum                                 -0.50899750            0.2054
49899
##                              B.pseudocatenulatum    Acetate  Propionate
## Bifidobacterium..taxid.1678.           0.3085263 -0.1020584  0.18923325
## B.adolescentis                         0.1569343 -0.3012401  0.47228313
## B.bifidum                              0.3453953  0.2102477 -0.06454972
##                                 Butyrate Isobutyric.acid Isovaleric.aci
d
## Bifidobacterium..taxid.1678.  0.09594782      0.09413161      -0.260759
5
## B.adolescentis                0.10080137      0.22335082      -0.341962
7
## B.bifidum                    -0.14356319     -0.52877874      -0.327196
1
##                              Valeric.acid
## Bifidobacterium..taxid.1678.   -0.1841357
## B.adolescentis                 -0.1326516
## B.bifidum                       0.0000000
```

```
head(corr_Bifido.red_SCFA$R$P, 3)
```

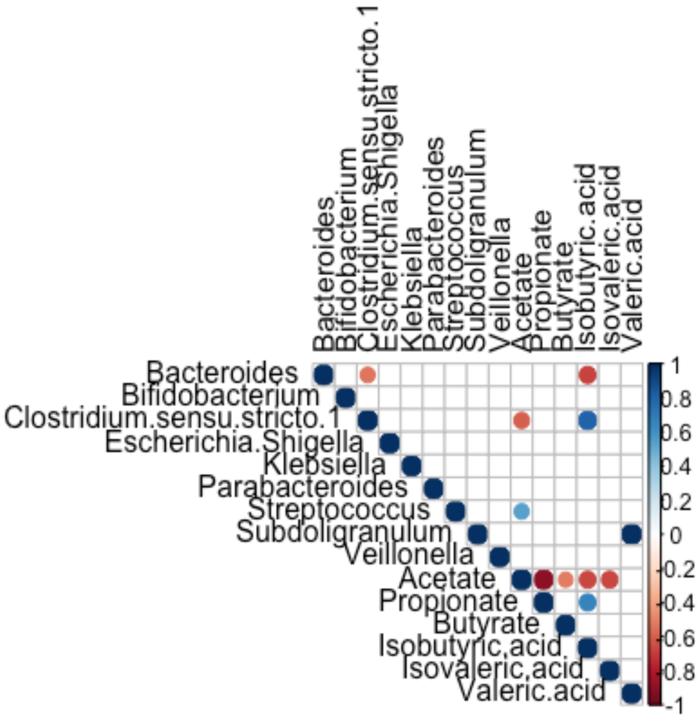
```
##                              Bifidobacterium..taxid.1678. B.adolescenti
s
## Bifidobacterium..taxid.1678.                           NA      0.315509
6
## B.adolescentis                                  0.3155096             N
A
## B.bifidum                                       0.5233444      0.929394
7
##                              B.bifidum   B.breve B.kashiwanohense  B.lo
ngum
## Bifidobacterium..taxid.1678. 0.5233444 0.1854036        0.5079134 0.727
8233
## B.adolescentis               0.9293947 0.6847972        0.8893619 0.913
3867
## B.bifidum                           NA 0.8367105        0.5493899 0.372
7746
##                              B.longum.subsp..infantis B.longum.subsp..l
ongum
## Bifidobacterium..taxid.1678.               0.77390255              0.86
21064
## B.adolescentis                             0.93499082              0.97
83115
## B.bifidum                                  0.05265065              0.46
26041
##                              B.pseudocatenulatum   Acetate Propionate
Butyrate
## Bifidobacterium..taxid.1678.           0.2632147 0.7174071 0.49938418 0
.7337420
## B.adolescentis                         0.5764656 0.2752382 0.07546689 0
```

```
.7207584
## B.bifidum                              0.2073427 0.4519842 0.81921976 0
.6097443
##                                Isobutyric.acid Isovaleric.acid Valeric.ac
id
## Bifidobacterium..taxid.1678.      0.73861793       0.3478928     0.51121
98
## B.adolescentis                    0.42360959       0.2121951     0.63744
15
## B.bifidum                         0.04270087       0.2338861     1.00000
00
```

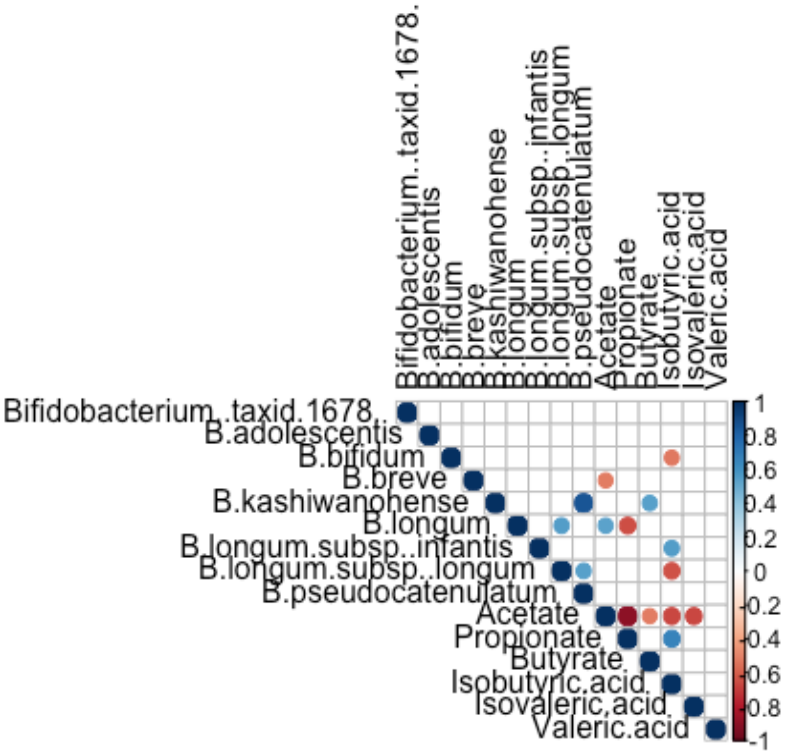
The correlation plots used the p-values in the rho-matrix and measured up against the significant level. The type was set to only show an upper triangular matrix. p.mat is the matrix of p-values that was considered against the significant level, which was set to 0.05. The insig shows the specialized insignificant correlation coefficients, and with blank, the corresponding symbols will not be visible. tl.col is the color of the text labels, which was set to be black.

```
library(corrplot)

corrplot(corr_16S_SCFA$R$r,
         type = "upper",
         p.mat = corr_16S_SCFA$R$P,
         sig.level = 0.05,
         insig = "blank",
         tl.col = "black")
```



```r
corrplot(corr_Bifido.red_SCFA$R$r,
         type = "upper",
         p.mat = corr_Bifido.red_SCFA$R$P,
         sig.level = 0.05,
         insig = "blank",
         tl.col = "black")
```



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

Postboks 5003
NO-1432 Ås
Norway