Norwegian University
of Life Sciences

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

Utilisation Potential of Human Milk Oligosaccharides and Mucin by *Ruminococcus gnavus* in the Human Infant Gut

Marte Bergene
MSc Biotechnology

Utilisation Potential of Human Milk Oligosaccharides and Mucin by *Ruminococcus gnavus* in the Human Infant Gut

Norwegian University of Life Sciences (NMBU),

Faculty of Chemistry, Biotechnology and Food Science

Acknowledgments

This thesis was performed at the Norwegian University of Life Sciences at the Faculty of Chemistry, Biotechnology and Food Science, under the supervision of Professor Knut Rudi (main) and PhD Morten Nilsen (co).

First, I would like to thank Knut Rudi for including me in the PreventADALL study, and for answering all questions and helping me throughout the writing process. Your dedication and knowledge are admirable. I would also like to thank Morten Nilsen for all the help with the experiment and for all data processing. Your knowledge and abilities have been much appreciated.

I would like to thank fellow master student Tonje Nilsen for the great collaboration. You made the laboratory work easier and more enjoyable. I would also like to thank the rest of the members of the Microbial Diversity (MiDiv) group for all their help and for making the laboratory environment fun, inspiring and inclusive, and a pleasure to be a part of.

Lastly, I would like to thank my family for always supporting me in everything I do. I would also like to thank my housemates for all the support, encouragement and good times.

Ås, May 2021

Marte Bergene

Abbreviations

ACN	Acetonitrile
AmBic	Ammonium bicarbonate
cDNA	Complementary deoxyribonucleic acid
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
ddNTP	Dideoxyribonucleotide triphosphate
DTT	Dithiothreitol
FID	Flame ionisation detector
Gal	Galactose
GalNAc	N-acetylgalactosamine
Glc	Glucose
GlcNAc	N-acetylglucosamine
GC	Gas chromatography
GH	Glycoside hydrolase
HMO	Human milk oligosaccharide
HPLC	High-performance liquid chromatography
IBD	Inflammatory bowel disease
IBS	Irritable bowel syndrome
IAA	Iodoacetamide
IT-sialidase	Intramolecular sialidase
KEGG	Kyoto encyclopedia of genes and genomes
LC	Liquid chromatography
LNB	Lacto-N-biose
LNT	Lacto-N-tetraose
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
mRNA	Messenger ribonucleic acid
NA	Nucleic acid
Neu5Ac	N-acetylneuraminic acid
NGS	Next generation sequencing
OTU	Operational taxonomic unit
PCR	Polymerase chain reaction
PreventADALL	Preventing atopic dermatitis and allergies
qPCR	Quantitative polymerase chain reaction
QIIME	Quantitative insight into microbial ecology
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
SCFA	Short chain fatty acid
SDS	Sodium dodecyl sulfate
TBS	Tris-based saline
TFA	Trifluoroacetic acid
TOF	Time-of-flight

Abstract

The human gut microbiota plays an important role in the health and immune development of the body. Gut bacteria can utilise carbohydrates that are unavailable to human enzymes, like human milk oligosaccharides (HMOs) present in breast milk, and produce short chain fatty acids (SCFAs) as fermentation products. HMOs are thought to be a selective force for colonisation of the gut the first months of life. The gut microbe *Ruminococcus gnavus* is found present in both infants and adults, and has previously been associated with disorders like Crohn's and eczema, and general dysbiosis. *R. gnavus* is a known utiliser of mucin glycans produced by the epithelium of the intestine, which resembles HMOs in structure. The idea that *R. gnavus* might utilise HMOs instead of or in addition to mucin stirs the thought that colonisation of *R. gnavus* might affect the colonisation of *Bifidobacterium* and other favourable bacteria. The aim of this thesis was to investigate the mucin and HMO degrading potential of *R. gnavus* and the potential for propionic acid production by *R. gnavus* in the infant gut.

Faecal samples from 6-month-old infants, collected through the study Prevent Atopic Dermatitis and Allergies (PreventADALL), were analysed. Microbial composition was determined through 16S rRNA gene sequencing, short chain fatty acid composition was determined using gas chromatography, genes in the *R. gnavus* genome were identified through shotgun sequencing, proteins were identified using mass spectrometry, and expression of genes coding glycan degrading proteins was determined using quantitative PCR (qPCR).

The results showed a complete pathway for propionic acid production in the genome and proteome, indicating propionic acid production by *R. gnavus* in the infant gut, as previously shown in the adult gut. Fucosidases, β-galactosidases, sialidases and several mucin specific enzymes were identified in the proteome of all or some of the samples, while no HMO specific enzymes were found in any proteome. Complete pathways for degradation of glucose, galactose and N-acetylglucosamine (GlcNAc) were identified. The results indicate that mucin degradation is more important than HMO in *R. gnavus*, but that HMOs might be partially degraded. It is difficult to predict the preferred substrate, as many of the identified enzymes can be used on both mucin and HMOs. High abundance of *R. gnavus* is thought to be unfavourable, because of the properties of *R. gnavus* and the negative correlation with *Bifidobacterium*. Lack of *Bifidobacterium* is a sign of a more adult gut which is unwanted during infancy. The results from this thesis lay the foundation for further studies on glycan degradation by *R. gnavus*, like gene expression analysis and growth experiments on different mucins and HMOs.

Samandrag

Tarmmikrobiotaen i menneske spelar ei viktig rolle for kroppen si helse og immunutvikling. Tarmbakteriar kan nytta karbohydrat som ikkje er tilgjengelege for humane enzym, som oligosakkarid funne i morsmjølk (HMO), og produserer kortkjeda feittsyrer som fermenteringsprodukt. Ein trur HMOer har ei selektiv kraft ved kolonisering av tarmen dei fyrste levemånadene. Tarmmikroben *Ruminococcus gnavus* finnes i både spedborn og vaksne, og bakterien er tidlegare blitt assosiert med lidingar som Krohns, eksem og generelt ved dysbiose i tarmen. *R. gnavus* er kjend for å nytta mucin glykan produsert av tarmepitelet, som liknar HMO i struktur. Ideen om at *R. gnavus* kanskje nyttar HMO i staden for eller i tillegg til mucin har vekkja tankar om at *R. gnavus* kanskje påverkar koloniseringa av *Bifidobacterium* og andre gunstige bakteriar. Målet med oppgåva var å undersøkje potensialet *R. gnavus* har til å nytte mucin og HMO, og potensialet for produksjon av propionsyre i tarmen til spedborn.

Avføringsprøvar frå seks månader gamle spedborn, samla inn gjennom studien Prevent Atopic Dermatitis and Allergies (PreventADALL), vart analysert. Mikrobiell samansetjing vart bestemt gjennom 16S rRNA gensekvensering, samansetjing av kortkjeda feittsyrer vart bestemt ved bruk av gasskromatografi, det genetiske potensialet til *R. gnavus* vart identifisert ved shotgun sekvensering, protein vart identifisert ved massespektrometri og uttrykket av glykan-nedbrytande gen vart bestemt ved bruk av kvantitativ PCR (qPCR).

Resultata viste ein komplett produksjonsveg for propionsyre i genomet og proteomet, som indikerer produksjon av propionsyre frå *R. gnavus* i tarmen til spedborn, som tidlegare vist hjå vaksne. Fukosidaser, β-galaktosidaser, sialidaser og fleire mucin-spesifikke enzym vart identifisert i proteomet til alle eller nokre prøvar, men ingen HMO-spesifikke enzym vart funne. Komplette nedbrytingsvegar for glukose, galaktose og N-acetylglukosamine (GlcNAc) vart identifisert. Resultata indikerer at nedbryting av mucin er viktigare i *R. gnavus*, men at HMO kanskje blir delvis nedbrote. Det er vanskeleg å sjå kva substrat *R. gnavus* vil føretrekka, då *R.* gnavus kan nytte mange av enzyma på både mucin og HMO. Ein trur mykje *R. gnavus* i tarmen til spedborn ikkje er gunstig, på bakgrunn av eigenskapane til bakterien og den negative korrelasjonen til *Bifidobacterium* som er funne. Mangel på *Bifidobacteirum* er eit teikn på ein meir vaksen tarmmikrobiota, noko som ikkje er ynskja tidleg i livet. Resultata frå oppgåva legg grunnlaget for vidare studiar på glykan-nedbryting i *R. gnavus*, i form av genuttrykksanalyser og dyrkingseksperiment på ulike mucin og HMOer.

Table of contents

1. Introduction

1.1 The human gut microbiota

In the human gut, a microbial community live in a symbiotic relationship with humans. The microbial community consists of members of the domains Bacteria and Archaea and the kingdom Fungi, in addition to viruses and protists. This community is termed the human gut microbiota. The highest density of bacteria on the planet is found in the colon of humans, because of the human gut microbiota (Whitman et al., 1998). The normal adult gut microbiota consists of 150-200 bacterial species, with the most abundant phyla being Firmicutes and Bacteroidetes (Eckburg et al., 2005; Faith et al., 2013). Most bacteria are strictly anaerobic and beneficial for the host in some way. The microbiota can affect the immune system of the host and can induce inflammation and development of diseases and health problems. However, the first year of life the microbiota is important for development and maturation of the immune system. Bacteria in the gut can also give nutritional benefits and they can inhibit colonisation of pathogens. Bacteria are therefore important for human health and survival.

The gut microbiota can utilise nutrients unavailable to human epithelial cells, like the dietary fibres hemicellulose and resistant starch. The main products of bacterial carbohydrate metabolism are short chain fatty acids (SCFAs), which can be consumed by epithelial cells in the colon and can be used as energy source by the human body (Ganapathy et al., 2013). Vitamins, antimicrobials, and other compounds can be produced by the gut microbiota. The production of antimicrobial compounds by the gut microbiota prevents colonisation of opportunistic bacteria with harmful properties. Also, the presence of huge amounts of commensal bacteria in the colon prevent colonisation of opportunistic bacteria alone, as there are no room for new colonists. The human gut microbiota consists of commensal bacteria making nutrient accessible to the human body and protecting the human body from pathogens.

Colonisation of opportunistic bacteria and disturbance of the microbiota in the colon of humans can cause a variety of diseases and disorders. Gut microbes can produce cytokines, which induce production of immunoglobulin A in the epithelial cells, inducing T regulatory cells and an immune response (Geuking et al., 2014). Individual species can also cause diseases in the gastrointestinal (GI) tract of humans, such as *Vibrio cholerae* causing cholera

disease. Imbalance between the bacterial species in the gut plays a significant role in a lot of human gut disorders and can be caused as a result from use of antibiotics, change in diet or in combination with other diseases. The imbalance, called dysbiosis, is defined as deviations from a normal, healthy gut microbiota, termed normobiosis (Casén et al., 2015). Dysbiosis in the gut can lead to disorders like irritable bowel syndrome (IBS) and inflammatory bowel diseases (IBD), like Crohn's disease (Casén et al., 2015). Bacteria present in the human gut have also been associated with diseases and disorders like autism, Parkinson's disease, and diabetes (Bullich et al., 2019; Hughes et al., 2018; Qin et al., 2012). Although bacteria in the gut is a crucial part of the human body, imbalance in the bacterial community or colonisation of pathogens can be harmful.

Studying the microbiota is important to understand nutrient utilisation in humans, causes and development of diseases and disorders, maturation of the immune system, and the symbiosis of humans and microbes in general, as examples. A much studied field within gut microbiota is the colonisation process of neonates and infants. The effect of different parameters, like type of birth, feeding, health of mother, use of medicine and antibiotics, and environment, to name a few, is studied to try and identify differences in gut microbiota in infancy and to find associations with health conditions. These studies can help identify "good" microbes that lead to favourable development of the immune system and inhibition of pathogens, giving healthy and happy infants and adults.

1.2 Infant gut colonisation

The bacterial colonisation of the human gut is a gradual process, and whether colonisation of the gut starts before or after birth has not been established. Two hypothesis dominates today, where one claims the womb is not sterile and that foetuses are colonised before birth (Aagaard et al., 2014; Jiménez et al., 2005). The other, called the sterile womb paradigm, claims the foetus is first colonised when the foetal membrane ruptures (Lauder et al., 2016). Regardless of when the first colonisation happens, some of the first colonisers of the human gut are facultative anaerobic bacteria. Based on mode of delivery, vaginal or caesarean, the bacteria to colonise are believed to be present in the mother's vagina or gut, or present in the environment, respectively. Other factors like gestational age at birth, use of antibiotics, and diet (breastfeeding or formula) can also affect which species are early colonists (Milani et al., 2017).

Colonisation of new bacteria and the development of the microbiota in the human gut starts at birth and changes appear throughout life, but the microbiota is starting to stabilise and look similar to an adult microbiota at 2-5 years of age (Cheng et al., 2016; Ringel-Kulka et al., 2013). The microbiota of neonates has low α-diversity and consists mostly of facultatively anaerobic species belonging to phyla Actinobacteria and Proteobacteria (Milani et al., 2017). For instance, infants born vaginally will be exposed to bacteria dominant in the vagina, like *Lactobacillus* and *Prevotella*, and develop a microbiota dominated by these bacteria (Dominguez-Bello et al., 2010). Infants born with caesarean section will not be exposed to vaginally associated bacteria, and will develop a different type of microbiome, based on environmental and skin associated bacteria. In the case of diet, it is well known that breastfeeding support colonisation of *Bifidobacterium*, which utilises human milk oligosaccharides and is thought to be beneficial for the infant health and development, as it contributes to delayed colonisation of other bacteria. At one year the level of bacteria in the gut has increased and consist of strict anaerobic bacteria from phyla Firmicutes and Bacteroidetes (Avershina et al., 2016).

1.3 *Ruminococcus gnavus*

The gut microbe *Ruminococcus gnavus* has been found present in the gut microbiome of over 90% of adults and is thought to be an essential part of the microbiome of the gut (Qin et al., 2010). The species belongs to family *Lachnospiraceae* in phylum Firmicutes and contain obligate anaerobic and gram-positive bacteria (Moore et al., 1976). The genus *Ruminococcus* has been found to be dominant only a few days after birth, and *R. gnavus* has been found present in both breastfed infants and those not breastfed, in approximately equal amounts (Favier et al., 2002; Sagheddu et al., 2016). The bacteria utilize fermentable carbohydrates as energy and carbon source, and produce fermentation products like acetic acid, formic acid and ethanol (Moore et al., 1976). Nilsen et al. (2020) showed that *R. gnavus* was negatively correlated to the short chain fatty acid butyrate, while Crost et al. (2013) has shown propionate production from degradation of fucosylated sources by *R. gnavus*.

R. gnavus has been associated with several diseases and dysbiosis in the gut. IBD is caused by dysbiosis in the gut, and *R. gnavus* has previously been associated with IBD. Transient increased abundance of *R. gnavus* has been associated with active periods of disease (Casén et al., 2015; Hall et al., 2017). *R. gnavus* has also been shown to express high β-glucuronidase

activity, which lead to inflammation, and to produce glucorhamnan, a polysaccharide which directly can induce inflammation (Henke et al., 2019; Joossens et al., 2011). High abundance of *R. gnavus* in the adult gut has also been associated with eczema and generalised anxiety (Jiang et al., 2018; Zheng et al., 2016). High abundance of *R. gnavus* seems to increase the risk of inflammation and gut related diseases in the adult gut.

1.3.1 Mucus utilisation by *R. gnavus*

R. gnavus is known for its ability to utilize host derived mucin glycans as carbon source (Crost et al., 2013). Mucin glycans are O-linked glycoproteins with α- and β-linked N-acetylgalactosamine (GalNAc), galactose (Gal) and N-acetylglucosamine (GlcNAc), produced by goblet cells in the epithelium. The structure of the glycans can be elongated and modified with α-1,2/3/4-linked fucose and α-2,3/6-linked sialic acid. The mucus layer of the small intestine consists of one layer, while two layers are present in the colon. The inner mucosal layer of the colon is impermeable to microbes, giving the mucosa a protective function and separating the bacteria from immune cells and epithelial cells, while the outer mucosal layer is the habitat of some commensal bacteria, like *Lactobacilli* and *Ruminococcus* (Johansson et al., 2011).

The strain *R. gnavus* ATCC 29149 possesses the ability to release 2,7-anhydro-Neu5Ac from sialylated substrates, like mucus, by the action of an intramolecular trans-sialidase (IT-sialidase)(Crost et al., 2016). Most other mucin-degrading bacteria in the gut release the sialic acid N-acetylneuraminic acid (Neu5Ac) from sialylated substrates, and Crost et al. (2016) speculated that the IT-sialidase might give *R. gnavus* and advantage in mucus-degradation. The IT-sialidase is part of a gene cluster called the *nan*-cluster (Bell et al., 2019). The cluster includes genes coding N-acetylmannosamine-6-phosphate 2-epimerase (*nanE*), a sialidase (*nanH*), ABC-transporter components (*nanT*), a N-acetylmannosamine kinase (*nanK*) and a Neu5Ac lyase (*nanA*), and are used to metabolise and transport sialic acid (Crost et al., 2013). The *nan*-cluster is present in several bacteria capable of using mucus as sole carbon source. *R. gnavus* is also shown to utilise fucose from host glycans, by use of fucosidases coded in the genome (Crost et al., 2013).

1.4 Human milk oligosaccharides

Breast milk of humans differs from breast milk of other mammals, as it contains more complex and diverse oligosaccharides in higher concentrations (Bode, 2012). Normally, breast milk functions primarily as nutrient, but human breast milk acts in addition as protection against pathogenic colonisation and infection (Urashima et al., 2001). The composition of human milk oligosaccharides (HMO) in breast milk differs slightly between individuals. The types of fucosylated HMOs present in breast milk is determined by the secretor and Lewis blood group genes and the expression of these in the mammary glands (Blank et al., 2012). The genes code fucosyltransferases used to produce Lewis and ABO antigens. Fucosyltransferases can also be used to fucosylate HMOs, and depending on what genes are expressed, different proteins are produced which can make α-1,2- or α-1,3/4-linkages with fucose. Non-secretor individuals who also are Lewis blood group negative will not produce fucosyltransferases, and no fucosylated oligosaccharides are secreted. Because of difference in gene expression, the expression of HMO patterns is thought to be individual.

Human milk oligosaccharides are complex carbohydrates with a core made of the three monosaccharides glucose, galactose and GlcNAc, and can be extended with fucose and/or Neu5Ac (Figure 1.1)(Bode, 2012). Fucose can be linked to both glucose (α-1,3 linkage), galactose (α-1,2 linkage) and GlcNAc (α-1,3 or α-1,4 linkage), while Neu5Ac can be linked to galactose (α-2,3 or α-2,6 linkage) and GlcNAc (α-2,6 linkage)(Zúñiga et al., 2018). The most common HMOs present in human breast milk are lacto-N-tetraose and lacto-N-neotetraose, both with and without fucose and Neu5Ac extensions. The structure of HMOs resembles that of mucin, glycoproteins present at the mucus surface of the intestine, as previously described (paragraph 1.1.3.1).

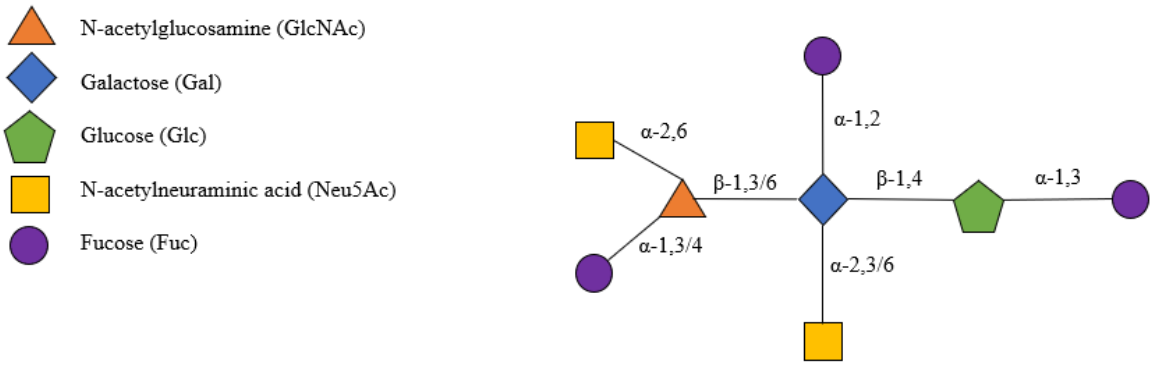

Figure 1.1. Structure of HMO. *Simplified figure of building blocks in HMOs and the type of chemical bonds they can form with each other.*

There is thought to be a microbe-host coevolution between the gut microbiota and oligosaccharides present in human breast milk, as HMOs contribute to shape the human gut microbiota through the first months of life. Gut microbes have adapted genetically to the host glycans by harbouring genes coding glycosyl hydrolases (GH) and galactosidases, giving an advantage in early colonisation (Milani et al., 2017). GH can break up the structure of HMO and utilise the energy hidden in the core structure, as well as utilising fucose and sialic acid. It is discussed if the sialylation of HMOs contributes to brain development, as early brain development requires high levels of sialic acid as brain nutrient. Levels of sialic acid in the brain has been measured higher in breast fed infants than in formula fed infants, indicating that HMOs might be a source of sialic acid (Wang et al., 2003). HMOs can also alter the immune system by influencing the lymphocytes and affect the production of cytokines, which is thought to alter the T-cell response (Eiwegger et al., 2004).

1.4.1 Utilization of HMO in the gut

The small intestine of humans does not seem to harbour the enzymes necessary to degrade HMOs, and the HMOs will therefor pass on to the large intestine where the microbiome utilises the glycans as energy source (Engfer et al., 2000). As the composition of HMOs differs between individuals, the gut microbiota composition of infants can differ slightly depending on the HMO degrading abilities of the microbes. Different bacteria have developed their enzymes to specifically target different chemical bonds in HMOs and utilises HMOs as one of, or the only carbon source (Garrido et al., 2015). Some bacteria are known to degrade HMOs, like members of the genera *Bifidobacterium* and *Bacteroides*, as these genera are

shown to be more present in breast fed infants and are shown to harbour GHs (Marcobal & Sonnenburg, 2012). A few bacterial species, like *Bifidobacterium infantis*, harbours the entire apparatus for HMO degradation and are early colonists of the infant gut. Because of the phenomenon of cross-feeding, the entire machinery is not necessary, and there are several bacterial species who can utilise only parts of HMOs (Milani et al., 2017).

1.5 Short Chain Fatty Acids

Non-digestible carbohydrates, like cellulose, resistant starch, lignin and pectin, as well as human milk oligosaccharides, remain intact in the gastrointestinal tract of humans until they reach the colon, as human enzymes cannot break the β-1-4-glycosidic linkage present in dietary fibres. Bacteria in the colon have enzymes with this ability and utilise dietary fibres and other complex carbohydrates, like HMOs, as carbon and/or energy source. The products of the carbohydrate degradation in gut bacteria are SCFAs, also called volatile fatty acids. Acetic acid, butyric acid and propionic acid are the most common SCFAs, which can be used as signalling molecules between colonic bacteria and the host, and as energy source for the body (Ganapathy et al., 2013).

Butyric acid is the most important SCFA and contribute to cell differentiation, apoptosis of cancer cells and inhibition of inflammation (Ganapathy et al., 2013). Lack of butyric acid in the colon could lead to autophagy of the epithelial cells, where the cells degrade own cell material, as butyric acid is the primary energy source of the epithelium (Donohoe et al., 2011). Autophagy of colonocytes is critical, as it can lead to inflammation and damage on the epithelial wall of the colon. Abundance of SCFAs, and particularly butyric acid, can regulate the permeability of the epithelium layer by regulation of proteins in tight junctions (Morrison & Preston, 2016). Lack of butyric acid can increase the permeability, leading to transport of bacteria or bacterial components through the epithelium, which can cause inflammation. Propionic and acetic acid are absorbed by the epithelium and most of the acids are transported to the liver, where they are part of gluconeogenesis (Wong et al., 2006). Acetic acid can also be absorbed in the muscles and be used for lipogenesis.

1.5.1 Production and consumption of SCFA in the gut

The amount of SCFA in the colon changes during the first year of life, in accordance with changes in the composition of the gut microbiota (Nilsen et al., 2020; Tsukuda et al., 2021). In both infants and adults, acetic acid is the dominant SCFA, but the percentage of acetic acid in the gut is reduced from 3 to 12 months of age. The percentage of both butyric and propionic acid increases in the same period, and at 12 months the relative abundance of the SCFAs starts resembling that found in adults (Nilsen et al., 2020). The average molar ratio of acetic, propionic and butyric acid in adults is considered to be 60:20:20, respectively (Wong et al., 2006). Butyric acid producing bacteria, like *Faecalibacterium prausnitzii*, *Eubacterium rectale* and *Roseburia*, and propionic acid producing bacteria, like *Blautia* and *Roseburia*, belong to order Clostridiales (Louis & Flint, 2017). *Bifidobacterium* has been found as acetic acid producer, and the amount of Bifidobacteria present in the gut decreases from 6 to 24 months of life (Tsukuda et al., 2021).

SCFAs is not only used by the host, but can also be used by other bacteria, through cross-feeding (also called syntropy). Fermentation product from one bacterium can be used as substrate for another bacterium, generating other fermentation products. It is shown that acetic acid is consumed by butyric acid producing bacteria, resulting in interconversion of acetic to butyric acid in the gut (Barcenilla et al., 2000). Some interconversion from butyric to propionic acid was also shown (Besten et al., 2013). A snapshot of the SCFA levels in the gut, as in a sample, does therefore not show the total production and consumption of SCFA, but the net production. Cross-feeding enables more diversity in the microbiome, as bacteria can use both macromolecules from the diet and products from bacterial fermentation as energy source.

1.6 Gas chromatography and analysis of short chain fatty acids

Gas chromatography (GC) is one of the most used analytic tools in chemistry, used to separate and detect volatile organic molecules, like short chain fatty acids, or gasses (Linde.AG, 2021b). A GC is composed of an autosampler, an inlet, a column, a detector and a computer. The samples are injected into the inlet by an autosampler, where the samples are mixed with a carrier gas (mobile phase). In the inlet the sample is vaporised, if not in the gas phase. The vaporised sample is transferred to the column (stationary phase), where the

molecules are separated based on interactions with the stationary phase. When the molecules reach the end of the column, they are detected, and the computer generates a chromatogram. Based on the chromatogram the molecules present in the sample can be identified and quantified. To get more extensive information, a mass spectrometer can be used.

The mobile phase of the GC is a carrier gas transporting the vaporised sample through the column (Linde.AG, 2021b). It is important that the carrier gas does not react with the stationary phase in the column, and inert gasses are therefore often used, such as nitrogen, or helium or hydrogen gas. The stationary phase, the column, is covered with a liquid or film on the inside, which interacts with the molecules of the sample based on structure. An example is the stationary phase of polyethylene glycol (PEG), which is a good option to separate molecules containing hydrogen bonds, like acids and alcohols. As there are different types of columns, there are different types of detectors. Some of the most common detectors are flame ionisation detectors (FID), electron capture detectors (ECD) and flame photometric detectors (FPD). FID responds to C-H bonds and can detect hydrocarbons and other volatile organic compounds (Linde.AG, 2021a). In the FID detector there is a combustion of the sample, generating ions and free electrons. The ions and electrons wander in an electric field in the detector, and the flow of charged particles are detected. Choosing the right mobile and stationary phase and detector is important to obtain as good a result as possible for the specific sample.

1.7 Molecular Methods

Gut bacteria are difficult to study *in vitro*, as they are difficult to cultivate. Prediction of optimal conditions and nutritional needs is challenging, and depending on the sample, the amount of bacteria present can be limiting. Cultivation of bacteria living in complex microbial communities can also be difficult, as they may be dependent on other species for survival, through protection and cross-feeding. The evolution of metaomics; metagenomics, metatranscriptomics and metaproteomics, have made the study of bacteria much easier, as no cultivation is needed. An organism's genome contains the recipe for all proteins and cell components and reflects the metabolic potential of the organism. To determine what the organism could be doing at a given moment in an environment, one can study the transcriptome of the organism. The transcriptome reflects what genes are turned on at that exact moment, and up and down regulation of genes. It is not said that all messenger

ribonucleic acids (mRNAs) are translated to proteins, and to tell exactly which proteins are being used in the cell, the proteome can be studied. The proteins present in the cell will tell the exact function of the genes expressed, as post translational modifications cannot be predicted by studying the genome or transcriptome. Exactly how the organism is behaving in their natural habitat can only be determined with *in vivo* experiments, which is a downside to metaomics. As *in vivo* experiments can be difficult to conduct, in *vitro* experiments, like cultivation, can give an indication of how the organisms are behaving in their natural environments.

1.7.1 Nucleic acid extraction

To access the genome and transcriptome of the cells, the nucleic acids (NAs) deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) must be release from the cells, by cell lysis. Cell lysis can be induced mechanically, physically, chemically, or biologically. Chemical and biological lysis uses compounds like detergents and enzymes, respectively, to break down the cell wall and cell membrane. Heat or osmotic shock can be used to induce cell lysis physically, by altering the shape and rigidity of the cell wall and cell membrane. Heat can also destroy the DNA/RNA in the sample and must therefore be used with ease. Mechanical lysis opens the cells by beating, using beads or high pressure. Bead beating will destroy the cell wall and cell membrane, and the DNA and RNA will be released to the solution.

Prior to analysis, the NAs must be isolated. A much-used method for isolation of DNA and RNA is by magnetic silica particles. In presence of certain salts, a cationic bridge is formed between the particles and NAs (Boom et al., 1990). A magnet can be used to retain the particles in the tube and remove the solution containing cell debris and other components of the sample, and wash the nucleic acids with ethanol/buffer. As NAs are bound reversely to the particles they can be eluted with water or elution buffer. Isolated RNA must be treated with DNase to ensure no DNA is present in the sample. To further study the RNA, it must be transcribed into complementary DNA (cDNA), as RNAs are very unstable and only DNA can be amplified using polymerase chain reaction (PCR).

1.7.2 Polymerase Chain Reaction

PCR is a technique used to amplify fragments of DNA. Thousands of copies can be made in a short time, which makes the study of DNA much easier. The method was invented in 1985 and was based on the enzyme Taq DNA polymerase from *Thermus aquaticus*. The Taq DNA polymerase is heat stable and can withstand high temperatures and temperature cycling. A polymerase chain reaction consists of x numbers of cycles of denaturing, annealing and elongation. The DNA is denatured at about 95°C, generating single strands. During annealing specific primers are bound to single stranded DNA at about 55-60°C, before the sequence is elongated by DNA polymerase at about 72°C.

PCR can be qualitative, amplifying and detecting DNA fragments, or quantitative, measuring DNA concentration. Qualitative PCR is also called end point PCR, meaning the presence of DNA fragments is detected after the PCR, often using gel electrophoresis. Quantitative PCR is also called real-time PCR, as signals from DNA fragments are collected after each cycle of the polymerase chain reaction. Fluorogenic compounds give of fluorescence signals that are registered by the machine. One type of fluorogenic compound, DNA binding dyes, will give of signal when bound to double stranded DNA, during elongation. Another type of compound consists of a probe complementary to a sequence on the target fragment. The probe binds to single stranded DNA and give of fluorescence signal when the probe is broken up by the DNA polymerase. Quantification of PCR fragments is done measuring the amount of fluorogenic signals in each cycle.

1.7.3 Sequencing technologies

Determining the base sequence of DNA, called sequencing, has simplified the identification of bacteria, and made it possible to study genes and gene functions. Sequencing became a phenomenon in 1977 with the introduction of Sanger sequencing technology, which was the most used sequencing technology for 40 years (Sanger et al., 1977). Sanger sequencing uses deoxyribonucleotide triphosphate (dNTP) and end terminating dideoxyribonucleotide triphosphate (ddNTP) to copy DNA fragments, which results in fragments of varying length. The fragments are separated based on length using gel electrophoresis, and ddNTPs are identified by a bound fluorescens molecule. Based on the different lengths of the fragments, the ddNTP-signals are assembled into the DNA sequence of the template. Sanger sequencing

is also called first generation sequencing, in contrast to second generation sequencing, or next generation sequencing (NGS), which were introduced in 2005.

Next generation sequencing enables massive parallel sequencing of millions of short DNA fragments (up to ~400 bp)(van Dijk et al., 2014). A DNA sequencing library must be generated by amplification of the desired sequence and ligation of adapters and indexes to the sequences. Adapters enable the sequences to bind the flow cells, while the indexes make it possible to differentiate between the different samples being sequenced simultaneously. The sequences are bound to a flow cell and amplified, generating clusters, and are determined through base-calling, where each base emits a unique light signal registered by a computer. Next generation sequencing made sequencing easier and more effective, as multiple samples could be sequenced at the same time. It also led to the generation of huge amounts of data, which is the bottleneck of today's research. In recent years new sequencing technologies has emerged, called third generation sequencing. Third generation sequencing, also called single-molecule sequencing, is not dependent on cluster generation and can be done *in situ* (Schadt et al., 2010).

16S rRNA gene sequencing is one of the most common sequencing techniques for determining bacterial community composition. Earlier classification and identification of bacteria were based on morphology and metabolic properties. Sequencing of the 16S rRNA gene made it easier to classify bacteria, and the classification became more consistent and precise. The 16S rRNA gene is only present in prokaryotes and codes for the 16S subunit of the ribosome, which differs between bacterial species (Woese & Fox, 1977). Variable regions of the 16S rRNA gene makes it possible to distinguish between species, and sometimes even strains, while conserved regions make it possible to amplify. Often, only few variable regions are enough to differentiate between bacteria. The variable regions V3 and V4, for instance, are about 450 base pairs and can easily be sequenced and does not need assembling (Vargas-Albores et al., 2017). Different bacterial species are distinguished based on the identity of the 16S rRNA gene sequence and grouped in operational taxonomic units (OTUs). Bacteria with >97% sequence identity are often grouped together and are assumed to belong to the same bacterial species.

Shotgun sequencing is also used to determine the composition of bacterial communities, but it can in addition give information about bacterial functions in the community. In the shotgun

sequencing approach, the metagenome of a sample is fragmented randomly, and the random fragments are sequenced and assembled. The method can be used to assemble whole genomes and identify genes present in the metagenome. Shotgun sequencing generates huge amounts of data, which can be hard to process, but the information is highly valuable.

There are many different sequencing technologies available today, one of them being the Illumina sequencing technology. Illumina uses sequencing-by-synthesis, generating unique signals as the enzyme copies the template. DNA fragments are added index-sequences, containing adapters and primer sequences. The primer sequences are necessary for cluster generation, where clusters are generated through bridge amplification on the flow cell. The cluster density is important for optimal sequencing, as too low density will lead to poor fluorescence signals, while too high density will lead to difficulties distinguishing the fluorescence signals from different clusters. The amount of DNA added to the flow cell are therefore crucial to get a successful sequencing.

1.8 Protein analysis using mass spectrometry

Proteomics is the study of all proteins present in a system, and it is a useful tool to fully understand how a biological system works. The proteome of a cell gives an insight to the actions of the cells from the exact moment the sample is collected. By breaking down the protein into its amino acid units, the protein can be identified and post translational modifications can be found. These modifications will affect the function of the protein and cannot be determined or anticipated from the DNA or the mRNA sequence. One protein may have several functions, as well as several different proteins may have the same function. There is no coherence between the level of protein in a cell and the level of mRNA coding the protein (Gygi et al., 1999), and proteomics can be used to decide what is in fact translated. The most used method for identification and quantification of proteins is mass spectrometry (MS)(Cravatt et al., 2007). MS can be used to analyse proteins as intact entities (top-down proteomics) or as fragmented peptides (bottom-up proteomics), by deducing the ion mass of the molecules (Aebersold & Mann, 2016).

Protein analysis starts by extraction and isolation of proteins. Extraction of proteins from complex samples often involve several filtrations, to extract the bacterial cells, and cell lysis, to access the proteins inside the cells. When working with faecal samples, it is important to

filter out the eukaryotic cells, but retain the prokaryotic cells. A delicate filtration is therefore needed. The isolation of the extracted proteins is done using 2D gel electrophoresis, which alternatively can be used to purify/clean up an extracted metaproteome. In bottom-up proteomics proteins are fragmented by sequence specific enzymes, like trypsin, to peptides before analysis in the mass spectrometer (Aebersold & Mann, 2016). For further preparation of the peptides, they must be available. Transferring peptides from the gel to the liquid can be done by sonication, where the peptides are agitated by sound energy and are released from the gel. When the peptides are available, they can be purified for mass spectrometry using ZipTip pipettes (Merck Millipore, Cork, Ireland). The C18 material of ZipTips consist of hydrocarbon chains, which can bind and elute peptides. When peptides are bound to the C18 material, they can be washed before being eluted in a new solution.

The peptides in the sample can be analysed using liquid chromatography (LC) coupled to a MS. High-performance liquid chromatography (HPLC) is used to separate the components in the sample based on interaction created with a column material, resulting in different flow rates through the column. After separation, the peptides must be vaporised and ionised before identification. This can be done by electrospray ionisation (ESI), where liquid sample are sprayed and ionised in an electric field, or by matrix-assisted laser desorption ionisation (MALDI), where dry sample are ionised from a surface. Ionised peptides are transferred through a vacuum, where the mass of the ions is identified based on time-of-flight (TOF). To get a higher resolution, the peptides can be analysed using tandem mass spectrometry (MS/MS). After the peptides are vaporised and ionised, they pass through one mass analyser, before being fragmented by collision-induced dissociation (CID) using argon gas. The fragments are then passed through a second mass analyser (Lesk, 2016, p. 393-398). Time-of-flight, or velocity, is proportional to the mass-to-charge (m/z) ratio and is used to generate a MS/MS spectrum. From the MS/MS spectrum the amino acids and peptides can be identified by a computer. The intensity of the peaks in the spectra indicates the amount of the ion in the sample, and the peptides can be quantified.

1.9 The PreventADALL study

Faecal samples were obtained through the Prevent Atopic Dermatitis and ALLergies-study (PreventADALL). The study aims to understand and prevent development of allergies and atopic dermatitis in children. A total of 2397 mother-child pairs from Norway and Sweden were recruited. Through the study biological samples, including faeces and skin samples, were collected from mothers at 18 weeks pregnant and from children at 0, 3, 6, 9, 12, 24 and 36 months of age. Follow-up studies of the children are performed regularly. Data including information about external factors like type of birth, amount of breastfeeding, diet, weight, and health of parents was collected in addition to biological samples. Regular sampling from the children throughout the first years of life provides a unique chance to study the development of the microbiota and the effect on health conditions.

1.10 Aim of thesis

R. gnavus is found abundant in both infants and adults and colonises the gut early in life. In adults, *R. gnavus* is associated with mucus degradation, and can utilise mucin as a carbon and energy source. The structure of human milk oligosaccharides is resembling that of mucin, and the question has arisen if *R. gnavus* can utilise human milk oligosaccharides, as *R. gnavus* is abundant in infants. *R. gnavus* is also found to be a propionic acid producer when grown on mucin and fucosylated substrates. The main aim of this thesis was therefore to investigate the mucin and/or HMO utilisation properties of *R. gnavus* in the infant gut, and whether *R. gnavus* can contribute to propionate production in the infant gut. To achieve the main aim, following subgoals were studied:

- Identify gene coding potential HMO and mucin degradation proteins in the genome of *R. gnavus* present in the samples.
- Identify potential HMO degradation coding RNAs in the metatranscriptome.
- Identify proteins present in the bacterial cells in the samples.
- Analyse the presence of short chain fatty acids in the faeces.

2. Material and methods

A simplified overview of the study experiment is shown in Figure 2.1. A more extensive flowchart is found in figure A.1.

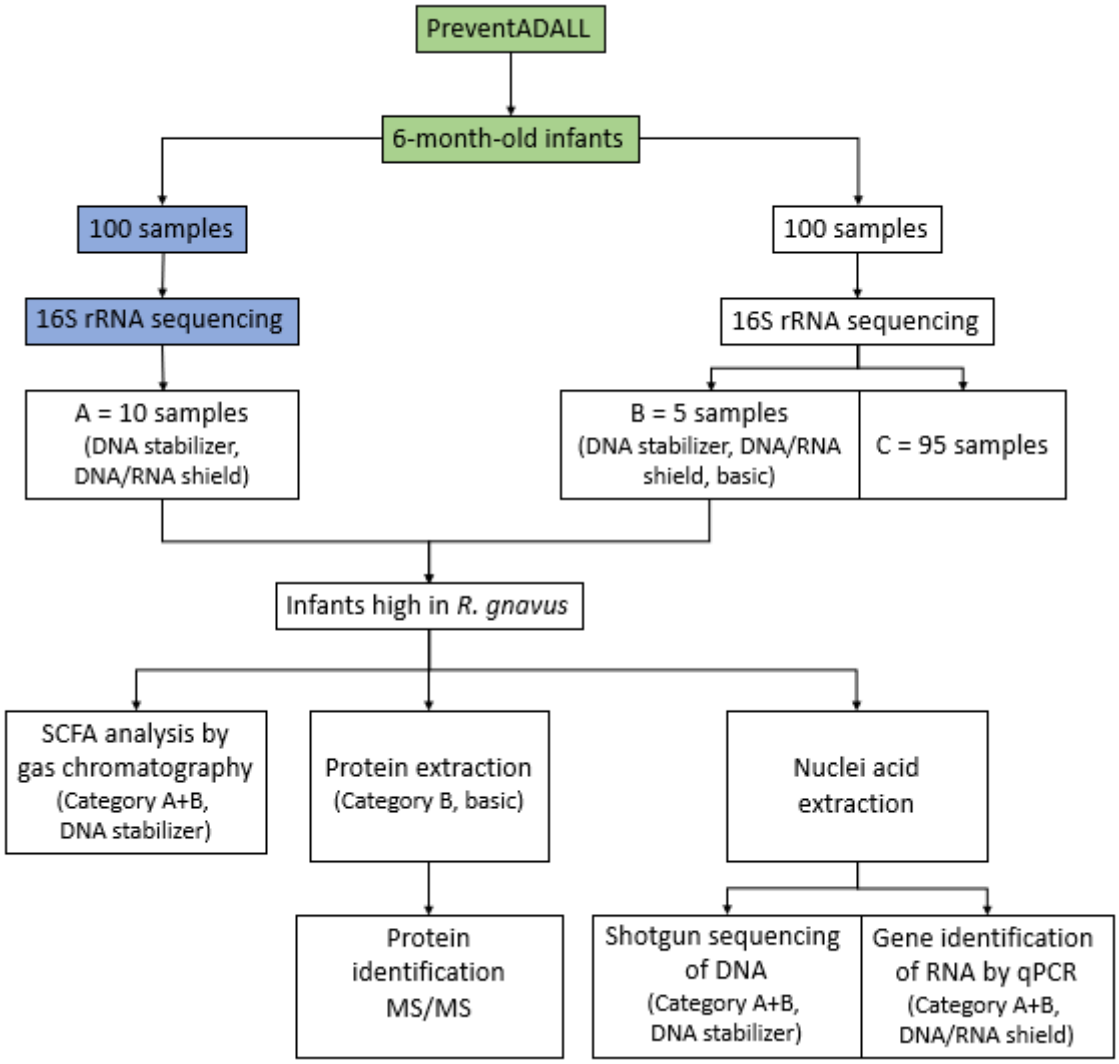

***Figure 2.1. Workflow overview of master thesis**. Faecal samples from 6-month-old children were previously collected by the PreventADALL study, and 15 samples were analysed in this thesis. The bacterial composition of hundred samples had previously been determined by 16S rRNA sequencing by PhD Morten Nilsen (blue boxes). Ten samples with high abundance of R. gnavus were chosen for further analysis, and grouped in category A. The ten samples could not be obtained without buffer (basic) and could not be used for protein analysis. The bacterial composition of 100 new samples were determined, and five samples with high abundance of R. gnavus were chosen for further analysis, and group in category B. The remaining 95 samples were grouped in category C and were not analysed further. Shotgun sequencing and SCFA analysis, by gas chromatography, were conducted on samples in category A and B. RNA gene identification by qPCR was conducted on category A and B, but one samples in category A was missing. Protein analysis using mass spectrometry was conducted on samples in category B. 16S rRNA gene sequencing results were analysed using QIIME, while shotgun sequencing results were analysed using a series of analysing tools. qPCR results were adjusted using LinReg, while mass spectrometry data were analysed using MaxQuant and Perseus. Correlation analysis was performed using RStudio, calculating Spearman correlations.*

2.1 The samples

Faecal samples obtained from 6-month-old infants through the PreventADALL-study were analysed using sequencing, gene identification, mass spectrometry and gas chromatography. The samples used for DNA sequencing was diluted 1:10 in DNA stabilizing buffer, while the samples used for RNA extraction was diluted 1:10 in RNA/DNA shield buffer. Samples used for protein extraction was stored without buffer (basic). All samples were stored at -80°C prior to analysis. A more detailed description of the experimental setup and the analyses is found in appendix A.

Hundred samples were used for 16S rRNA sequencing, where 53 of the infants were girls and 46 boys, and 16 were born with caesarean section. Of the 100 infants 62 were exclusively given breast milk at 3 months (48 exclusively breast fed) and 19 infants were given breast milk in addition to formula and/or solid food. At 6 months of age 72 infants were still given breast milk (breast fed and/or bottle fed). Of the 15 samples further analysed in the thesis, 11 infants were breast fed at 6 months, while one infant was not (three missing).

The samples analysed in this thesis were categorised into three groups based on the types of samples available for analysis. The 16S rRNA gene of 100 samples were sequenced, while 16S rRNA sequencing results from 10 addition samples were given. Of the 100 samples sequenced in this thesis, only five were used for further analysis. The samples were grouped into three categories: A (partial analysis) contained the 10 samples previously sequenced by PhD Morten Nilsen, which were not obtainable without buffer and could not be used for protein analysis, B (complete analysis) contained five samples available both with and without buffer and which were used for protein analysis, while C (no analysis) contained 95 samples which were not studied further.

2.2 Genomic DNA and RNA isolation and purification

2.2.1 Mechanical lysis

To isolate the bacterial DNA, the bacterial cells were lysed through mechanical bead beating using FastPrep96 (MP Biomedicals). The homogenised samples were centrifuged at 1200 rpm for 8 seconds, and 200 µL were transferred to a FastPrep tube (MP Biomedicals) containing 0.2 g acid-washed glass beads with size <106 µm and 0.2 g with size 425-600 µm together with 2 glass beads of size 2.5-3.5 µm (Sigma-Aldrich, USA). The samples were processed twice in FastPrep96 at 1800 rpm for 40 seconds with 5 minutes of rest between. The samples were centrifuged at 13 000 rpm for 5 minutes. The same procedure was used to lyse cells for RNA extraction.

2.2.2 DNA extraction

The mag midi DNA extraction kit (LGC Genomics, UK) was used according to the manufacturer's recommendations to extract DNA from the faecal samples. The kit uses paramagnetic particles that bind DNA because of its negative charge. In solutions with salts, salt bridges will be formed between DNA and the paramagnetic particles (Boom et al., 1990). Use of magnets makes it possible to remove all components not bound to the particles and isolate the DNA. Protease and lysis buffer is added to the samples. Protease will denature and break up proteins, making them easier to remove. Lysis buffer controls the viscosity and pH of the samples and will contribute to further lysis of the cell components remaining in the samples, by use of salts. The samples are washed twice, and DNA is eluted using elution buffer. Elution buffer disrupts the salt bridges between DNA and the particles, and DNA is eluted into the solution.

DNA from samples in category B and C were extracted using the ProteinaseLGCMini and MagMiniLGC procedures on a KingFisher Flex robot (Thermo Fisher Scientific, USA). DNA from samples in category A were extracted manually.

2.2.3 RNA extraction

For isolation of RNA molecules, the MagMAX^{TM}-96 Total RNA Isolation Kit (Thermo Fisher Scientific, USA) was used, following manufacturers recommendations. Phosphate

buffered saline (PBS)-washed *E. coli* DH5-α cells were used as positive control. Lysis/binding solution and bead mix were added to 30 µL of sample and incubated, to allow NAs to bind the paramagnetic beads. The NAs were washed with wash solutions using magnets, before TURBO DNase was added. The DNase solution will release the NAs from the beads and degrade the DNA present. RNA rebinding solution was added to rebind the RNA molecules to the beads, before further washing and elution. The RNA molecules were eluted in 35 µL elution buffer and treated with additional DNase, using the TURBO DNA-free™ Kit (Thermo Fisher Scientific, USA), as recommended by manufacturer. There was added 3.4 µL 10X TURBO DNase Buffer and 1 µL TURBO DNase, and the samples were incubated for 30 minutes. After incubation, 3.8 µL DNase inactivation reagent was added and 37 µL of the samples were transferred to a new tube at the end of the procedure.

2.2.4 cDNA synthesis

Extracted RNA was transcribed to cDNA by mixing 5 µL template RNA with 2µL 10x RT Reaction Premix with Random primers, 1.5 µL FIREScript Enzyme Mix and nuclease free water up to 20 µL per reaction, using the FIREScript RT cDNA synthesis mix kit with random primers (Solis BioDyne, Germany). Three parallels were made containing FIREScript enzyme mix and three parallels without the enzyme mix. The three identical parallels for each sample were mixed after the cDNA synthesis, to obtain as identical cDNA templates as possible during qPCR. The program for cDNA synthesis was as following: primer annealing at 25°C for 10 minutes, revers transcription at 50°C for 60 minutes and enzyme inactivation at 85°C for 5 minutes. The samples were kept at 4°C overnight.

2.3 DNA and RNA quantification

2.3.1 Qubit

Qubit fluorometer (Qubit 9V, Invitrogen, USA) was used to quantify DNA after extraction and after pooling of sequencing libraries. It was also used for quantification of RNA after DNase treatment. Fluorometers use fluorogenic dyes to identify genetic material in samples, and only excites signal when bound to target. The Qubit™ dsDNA HS Assay Kit (Thermo Fischer Scientific, USA) was used to quantify DNA, using manufacturers recommendations

and 2 µL template. For quantification of RNA, the Qubit™ RNA HS Assay Kit (Thermo Fischer Scientific, USA) was used, using the same approach as with DNA.

The Quant-iT™ Assay (Thermo Fischer Scientific, USA) was used to quantify indexed 16S amplicons using Cambrex – FLX 800 CSE (Thermo Fischer Scientific, USA). Two microlitre DNA template was mixed with 70 µL working solution (1:200). A few samples, ranging from lowest to highest Cambrex-value, was quantified using Qubit fluorometer, and was used to make a standard curve. The standard curve was used to convert the Cambrex-values to ng/µL.

2.3.2 Quantitative PCR

Concentration of extracted DNA used for 16S rRNA gene sequencing were checked by qPCR targeting the V3-V4 region of the 16S rRNA gene on a CFX96 Though (Bio-Rad, USA). HOT FIREPol EvaGreen qPCR supermix (1x, Solis BioDyne, Estonia), 0.2 µM forward primer and reverse PRK primer (table B.1)(Yu et al., 2005), 2 µL DNA-templat and nuclease free water were mixed to a final volume of 20 µL. Following program were used for amplification: 95°C for 15 minutes, followed by 40 cycles of denaturing at 95°C for 30 seconds, annealing at 59°C for 30 seconds and elongation at 72°C for 45 seconds. Melting curve analysis was added, with 5 seconds at each 0.5°C increase in temperature from 65°C to 95°C.

Identification of specific genes present in extracted RNA was performed by qPCR of cDNA using specific primers. HOT FIREPol EvaGreen qPCR supermix (1x, Solis BioDyne, Germany) was mixed with 0.2 µM of forward and revers primer, 2 µL cDNA-template and nuclease free water to a final volume of 20 µL. Primers targeting the 16S rRNA gene (PRK primers, table B.1) was applied to all samples and controls. The following primers were all applied to all samples for identification: RUMGNA_01058, RUMGNA_01638, RUMGNA_02693, RUMGNA_03611 and RUMGNA_03833 (table B.2)(Crost et al., 2013). The same program as described in the previous paragraph for qPCR analysis was used, but the annealing temperature was adjusted to 60°C.

The pooled 16S Illumina library was quantified using KAPA Library Quantification kit for Illumina Platforms (KK4824, Kapa Biosystems), according to the manufacturer's recommendations, using 2 µL template. Duplicates of each standard, and duplicates of 10^{-4} to

10^{-7} dilutions of the library were used. Quantification was done by initial denaturing at 95°C for 5 minutes, followed by 36 cycles of denaturing at 95°C for 30 seconds and annealing/extension at 60°C for 45 seconds. A melting curve analysis was added, as previously described (p. 17). The KAPA Library Quantification kit gives absolute quantification based on the oligo sequences present in adapter and on the flow cell, and not the 16S rRNA gene.

2.3.2.1 qPCR data processing

LinRegPCR version 2020.2 for analysis of real-time PCR data were used to determine baseline fluorescence and adjust the C_q-value.

2.4 Quality Assessment

2.4.1 Agarose gel electrophoresis

Gel electrophoresis was used to check the quality of DNA extraction, PCR products and sequencing libraries. Products in the shotgun sequencing approach were checked on 2% agarose gel at 80V for 45 minutes. The 16S rRNA gene sequencing products were checked on 1.5% Agarose gel at 80V for 30 minutes. Molecular Imager Gel DOC^{TM} XR Imagining Systems were used for visualisation of the gels.

2.5 DNA sequencing

2.5.1 16S rRNA gene sequencing

2.5.1.1 Amplicon PCR

The 16S rRNA gene was amplified through first step PCR. Master mix was made containing 1x HOT FIREPol Blend Master Mix ready to load (Solis BioDyne, Germany), 0.2 µM forward and reverse PRK primers (table B.1), 2 µL template DNA and nuclease-free water to a total volume of 25 µL. The fragments were amplified on a thermo cycler using the following program: 95°C for 15 minutes, followed by 30 cycles of denaturing at 95°C for 30 seconds, annealing at 55°C for 30 seconds and elongation at 72°C for 45 seconds. The amplification was ended by 7 minutes at 72°C and storage at 10°C.

2.5.1.2 Index PCR

Amplified 16S rRNA amplicons were indexed using 16 forward and 7 reverse index primers (table B.3). Indexes (5 µL, 0.2 µM) were distributed using the Eppendorf epMotion 5070 (Eppendorf AG, Germany). FIREPol Master Mix Ready to load (1x, Solis BioDyne, Germany), 2 µL template DNA and nuclease-free water was distributed to the indexes, to a total volume of 25 µL. The fragments were amplified using the following program: 95°C for 5 minutes, followed by 10 cycles of denaturing at 95°C for 30 seconds, annealing at 55°C for 1 minute and elongation at 72°C for 45 seconds. The amplification was followed by 7 minutes at 72°C and storage at 10°C.

2.5.1.3 Normalisation

Indexed 16S amplicon samples were normalised and combined to one library using Biomek 3000 (Beckman Coulter, USA). Volumes of each sample were calculated using the concentration from Cambrex and Qubit measurements. All volumes over 10 µL were downgraded to 10 µL, so only volumes between 1 µL and 10 µL were combined.

2.5.1.4 Clean-up of PCR products

The PCR products from first stage PCR of the 16S rRNA gene were purified using Sera-Mag beads (Sigma-Aldrich, USA) on Biomek 3000. Beads (1.0X) and 10 µL DNA-samples were used. The samples were washed with 80% ethanol and eluted with 20 µL nuclease-free water. The pooled 16S rRNA library was also purified using Sera-Mag beads, but the procedure was performed by hand, using 300 µL PCR product, 1.5x ampure beads and 40 µL nuclease-free water for elution. The concentration of PCR-products and the length of the fragments determine the concentration of beads used in the clean-up. Higher concentration of beads will bind shorter fragments.

2.5.1.5 Sequencing by Illumina MiSeq

The 16S amplicon library was sequenced using Illumina MiSeq (Illumina, USA). Before sequencing the pooled and normalised library was diluted to 4 nM using nuclease free water, before further dilution and denaturation following the protocol 16S Metagenomic Sequencing Library Preparation (Illumina, USA). The PhiX control was diluted using nuclease free water

instead of Tris. Both the library and the internal control PhiX was diluted to a concentration of 6 pM before combining the two to a final concentration of 20% PhiX and total volume of 600 µL.

2.5.1.6 Quantitative Insight Into Microbial Ecology (QIIME)

The data obtained after Illumina MiSeq sequencing of the 16S rRNA gene was processed using the Quantitative Insight Into Microbial Ecology (QIIME) pipeline. The data was first converted from a FASTQ file to a FASTA file, and the processing started by decomposing and filtering of poor-quality sequences (Huang, 2014). The barcodes were extracted, forward and revers reads were assembled, and the library was split into the respective samples. Reads were then grouped based on sequence identity, resulting in OTUs with over 97% sequence identity. Before grouping, the data was checked for chimeras, and there was set a cut-off on 5000 sequences per sample, meaning sequences read less than 5000 times during sequencing were removed. Using the SILVA database taxonomy was added to the OTUs using a consensus sequence from each OTU to search the database. Eventually, Shannon and Simpsons indexes for α-diversity and the Bray-Curtis dissimilarity index for β-diversity were calculated.

2.5.2 Shotgun sequencing

To prepare the samples for shotgun sequencing the Nextera DNA Flex Library Prep protocol was used, following Illumina's recommendation.

2.5.2.1 Tagmentation

The samples were tagmented using transposomes bound to paramagnetic particles, which both fragments the DNA and adds adapters to the fragments at once. The tagmented DNA fragments will remain bound to the beads. Beads containing transposomes was added to 30 µL DNA sample, and the tagmentation process was conducted at 55°C for 15 minutes. The tagmentation process was stopped by adding Tagment Stop Buffer and incubate the samples at 37°C for 15 minutes. Finally, the samples were washed three times with Tagment Wash Buffer.

2.5.2.2 Index PCR

Tagmented shotgun DNA was amplified and added indexes in one step. Enhanced PCR mix and i5 and i7 adapters were added to the beads with tagmented fragments (table B.4). The fragments were amplified using a thermal cycler, and the number of cycles were calculated for each sample separately, based on the DNA concentration measured by Qubit after DNA extraction. Dependent on concentration the fragments were amplified through six, eight or 12 cycles. The samples were treated at 68°C for 3 minutes, 98°C for 3 minutes, followed by x cycles of 98°C for 45 seconds, 62°C for 30 seconds and 68°C for 2 minutes, followed by 1 minute at 68°C and held at 10°C.

2.5.2.3 Clean up of library

The shotgun libraries were cleaned before pooling. From the amplified samples, 40 µL tagmented DNA was mixed with 72 µL Sample Purification Beads with ratio 1.8X. By use of magnet, supernatant was removed, and the samples were washed two times with 80% ethanol. Ethanol was removed and 32 µL Resuspension Buffer was added. The supernatants were transferred to a new plate and pooled.

2.5.2.4 Normalisation

For the ten samples in category A approximately equal amounts of DNA were added together in a pooled library, based on the samples with highest DNA concentration when measured by Qubit. The five samples of category B later sequenced varied more in concentrations and were pooled together with five samples to be sequenced by another master student. All samples therefore did not have the same concentration in the pooled library.

2.5.2.5 Sequencing by Illumina NovaSeq SP

Sequencing of the prepared library was done by Norwegian Sequencing Centre (NSC, Oslo, Norway). The library was sequenced using Illumina NovaSeq SP (Illumina, USA).

2.5.2.6 Processing of shotgun data

The quality of the shotgun sequencing raw data was checked through FastQC, giving both individual bases and whole sequences quality scores. The sequences were further processed

using several different tools, the first one being Trimmomatics (Bolger et al., 2014). Trimommatics filtered out sequences with poor quality scores and trimmed the ends of the sequences. Poor quality bases at the end of the reads were removed, as well as adapter sequences. Following parameters were used: MAXINFO: 50:0.24, Leading: 10, Trailing: 10, Slidingwindow: 5:20, Minlen: 32.

Some of the sequences sequenced were of human origin and had to be removed from the dataset. This was done using Bowtie2 and Samtools (Langmead & Salzberg, 2012; Li et al., 2009). The sequences were assembled into metagenomes using MetaSPADES (Nurk et al., 2017), which assemble through construction of deBruijn-graphs. Both MetaBAT2 and MaxBin (Kang et al., 2019; Wu et al., 2014) were used to make bins from the assembled metagenomes. Using dREP (Olm et al., 2017), the best bins from MaxBin and Metabat2, combined, were collected.

Bins were taxonomically classified using the Kraken2 standard Plus database (Wood & Salzberg, 2014), which classify bins using k-mers. Prodigal was used to collect amino acid sequences of the collected bins, and the amino acid sequences were annotated using CLC Genomic Workbench and InterProScan (Hyatt et al., 2010; Jones et al., 2014). To visualise possible pathways and proteins present in the metagenome, GhostKoala and Kyoto Encyclopedia of Genes and Genomes (KEGG) were used (Kanehisa et al., 2016).

The sequences were further filtered to make a database containing only amino acids sequences of DNA sequences mapping to *R. gnavus*. This was done using RStudio version 1.3.1093 (RStudioTeam, 2020). Bins belonging to *R. gnavus* was extracted based on DNA sequence. From the bins mapping to *R. gnavus*, amino acid sequences of contigs mapping to *R. gnavus* inside the bin were extracted. This resulted in a file containing all amino acid sequences of contigs mapping to *R gnavus* (figure C.1).

2.6 Protein analysis

2.6.1 Protein extraction and isolation

The faecal samples were fully suspended in 10 mL ice-cold TBS (tris-based saline) buffer. Two slightly different processes were used on the samples, called parallel 1 and parallel 2. Parallel 1 contained ~0.2g faecal sample, while parallel 2 contained 0.1-0.4g faecal sample. Parallel 1 was filtered over a 20 µm filter using Nylon-Net Steriflip® Vacuum filter unit (Merck Millipore, Fisher Scientific, USA) attached to a water suction pump, to remove large particles and most eukaryotic cells from the samples. The samples were centrifuged at 4000g for 10 minutes, to pellet the bacterial cells and smaller components, and the pellet was resuspended in 10 mL TBS buffer. Parallel 2 was not filtered over a 20 µm filter, but was centrifuged at 1500g for 5 minutes, and the supernatant was further centrifuged at 4000g for 10 minutes. The pellet was resuspended in 10 mL TBS buffer. Further processing was identical for both parallels. The suspension was filtered over a 0.22 µm nitrocellulose membrane filter (Millipore, USA) using a Millipore Vacuum Filtration System (Merck Millipore, USA) attached to a water suction pump. This filtration collected the bacterial cells on the filter, while small cell components and proteins present outside the cells were removed. The filter was cut and placed in FastPrep-tubes with 0.2g acid-washed glass beads (<106 µm), 0.2g acid-washed glass beads (425-600 µm), 2x 2.5-3.5 mm acid washed glass beads and 1 mL lysis buffer (50 mM Tris hydrochloride, 200 mM sodium chloride, 0.1% Triton-X100, 10 mM dithiothreitol, 2% sodium dodecyl sulfate (SDS)). Cells on the filter was lysed using FastPrep-96 (MP Biomedicals, USA) with 3x60 second pulses at 1800 rpm, followed by centrifugation at 16 000g for 15 minutes at 4°C. As much as possible of the supernatant was transferred to new tubes.

The amount of extracted protein was quantified using Pierce™ BCA Protein Assay Kit (Thermo Fisher Scientific, USA) on Eppendorf BioPhotometer® D30 (Eppendorf AG, Germany). The BCA kit uses the biuret reaction, where proteins are detected because of their reduction of Cu^{2+} to Cu^{+}. The quantification is based on the amount of Cu^{+} present. The kit is detergent-compatible, meaning that the reaction and detection tolerates relatively high concentrations of detergents, like SDS. A 1:5 dilution of the samples was measured, mixing 50 µL of the dilution with 1 mL of BCA working solution, made as recommended by manufacturer. A 1:5 dilution of lysis buffer was used as negative control/blank. The samples

were incubated at 60°C for 30 minutes and cooled to room temperature on ice before measuring the concentration at 562 nm.

Quantification of proteins was done to decide how much of the sample to apply to the gel in the next step. To isolate the proteins and remove compound that can affect further analysis, like SDS, the protein samples were run on a Mini-PROTEAN TGX stain-free gel (Bio-Rad Laboratories, USA). To the gel, 19.5 µL of the sample was added, together with 7.5 µL of sample buffer (Bio-Rad Laboratories, USA) and 3 µL of reducing agent (Bio-Rad Laboratories, USA). All samples with concentrations below 2.05 µg/µL were concentrated as much as possible before added to the gel. This was done using SpeedVac without heat. The reagents were mixed with sample and incubated at 90°C for 5 minutes, followed by centrifugation at 10 000g for 1 minute. The gel was run for about 7 minutes at 270V with Tris-Glycine-SDS (TGS) buffer. The gel was stained using 0.05% Coomassie in destaining solution (25% isopropanol and 10% glacial acetic acid) for 1 hour, and destained with destaining solution for 2x20 minutes and overnight in 1:2 dilution of destaining solution.

2.6.2 Protein purification and preparation

Preparation of protein samples were done using method developed by Arntzen et al. (2015). The coloured part of the gel was cut in 1x1 mm cubes and put in a tube with 200 µL MilliQ-water. The samples were incubated for 15 minutes on Eppendorf ThermoMixer (Eppendorf AG, Germany) at 22°C. The liquid was removed and 200 µL 50% acetonitrile (ACN)/25 mM ammonium bicarbonate (AmBic) was added. The samples were incubated for 15 minutes on thermo mixer at 22°C, before the liquid was removed. This was repeated once. Samples were incubated with 100 µL 100% ACN for 5 minutes on thermo mixer at 22°C. The liquid was removed, and the samples were air-dried for 1-2 minutes. ACN is a medium-polarity solvent that is used as mobile phase in HPLC and LC-MS, while AmBic is used in LC because of its buffer capacity.

Proteins in the samples were reduced by adding 50 µL dithiothreitol (DTT) solution (10 mM DTT/100 mM AmBic) and incubating for 30 minutes on thermo mixer at 56°C. DTT reduces the disulfide bonds between cysteines in proteins and will contribute to denature the proteins further. Samples were cooled down and DTT was removed, before adding 50 µL iodoacetamide (IAA) solution (55 mM IAA/100 mM AmBic). IAA will bind thiol groups in

cysteine and inhibit formation of new disulfide bonds. Samples were incubated for 30 minutes at room temperature in the dark, as IAA is light sensitive. IAA was removed, before adding 200 µL 100% ACN and incubating for 5 minutes on thermo mixer at 22°C. The liquid was removed, and the samples air-dried for 1-2 minutes.

To digest the proteins, 30 µL Trypsin solution (10 ng/µL trypsin/trypsin buffer) was added to the samples, and they were incubated for 30 minutes on ice. Trypsin breaks peptide bonds between lysine and arginine and will contribute to break down the proteins. Additional trypsin buffer (25 mM AmBic/10% ACN) was added to cover the gel pieces, before the samples were incubated at 37°C on thermo mixer overnight. Next day the samples were cooled down and added 40 µL 1% trifluoroacetic acid (TFA). TFA will inhibit trypsin and stop the protease from breaking down proteins further. The samples were stored in refrigerator for 20 days.

The samples were sonicated on water bath for 20 minutes, to transfer the proteins from the gel pieces to the liquid. The proteins were extracted from the liquid using ZipTips, containing a C18 material which can bind and elute proteins. Prior to isolation from each sample, the ZipTip must be conditioned and equilibrated to ready binding of proteins. 100% methanol (20 µL) was pipetted and discarded to waste to condition the pipette. Further conditioning was done by pipetting 20 µL 70% ACN/0.1% TFA and discarding to waste. Both methanol and ACN are organic solvents and are used for activation of the C18 material. 0.1% TFA (20 µL) was pipetted and discarded to equilibrate the C18 material, which will lower the pH of the material to enhance formation of ion bonds with the proteins. To bind the proteins to the C18 material, it was pipetted up and down four times in the sample. It was then pipetted and discarded 20 µL 0.1% TFA, to wash the proteins. The proteins were eluted in 20 µL 70% ACN/0.1% TFA, by pipetting up and down 4 times in the liquid. All liquid were removed using SpeedVac, before the samples were dissolved in 10 µL 2% ACN/0.1% TFA. The concentration of proteins was measured in 1.5 µL sample using Nanodrop (Thermo Scientific NanoDrop One Microvolume UV-Vis Spectrophotometer; Thermo Scientific, USA).

2.6.3 Protein identification by mass spectrometry

The isolated proteins were identified using nano-flow liquid chromatography mass spectrometry system (Dionex Ultimate 3000 UHPLC; Thermo Scientific, Bremen, Germany) connected to Q-Exactive mass spectrometer (Thermo Scientific, Bremen, Germany). The

proteins in the samples were loaded to a trap column and backflushed to an analytical column. In the analytical column the proteins were separated based on size and chemical properties. The proteins were ionised and transferred through a vacuum, where the TOF was measured, which was converted to ion mass. Some of the ions were then fragmented and passed through a second mass analyser. A more detailed description of the method and parameters, provided by PhD Morten Nilsen, is available in appendix D.

2.6.4 Processing of data from mass spectrometry

Raw files from MS were analysed and ions were identified using MaxQuant version 1.6.7.0, with the MaxLFQ algorithm for normalisation and intensity determination (Cox & Mann, 2008). The raw files were searched against the self-constructed database containing amino acid sequences from *R. gnavus.* It was also searched against the human genome (*Homo sapiens*, 73952 sequences) to remove contaminants. Common contaminants, as trypsin and human keratin, were complemented to the search in the database. Variable modifications were oxidation of methionine, conversion of glutamine to pyro-glutamic acid, protein N-terminal acetylation and deamination of glutamine and asparagine. Fixed modifications were carbamidomethylating of cysteine residues. It was also allowed with two missed cleavages of trypsin.

Perseus version 1.6.15.0 were used for filtering and other analysis of the protein data. Possible contaminants and human proteins were filtered away, and values were log2 transformed. Normalisation of results were checked studying the distribution of protein counts. Samples and proteins were clustered using hierarchical clustering. Clustering were based on Euclidean distance and average linkage.

2.7 Short chain fatty acid analysis using gas chromatography

Faecal samples in DNA stabilizing buffer were added 1:1 internal standard, containing 0.4% formic acid and 2000 µM 2-methylvaleric acid. Formic acid was used to activate the acids in the sample by lowering the pH, and the concentration of formic acid was too low to be detected by the detector. The 2-methylvaleric acid was added in know concentration to enable quantification of acids present in the samples. The samples were centrifuged at 13 000 rpm for 10 minutes to pellet large particles. The supernatant was centrifuged again at 10 000 rpm

for 5 minutes, using a filter column (0.2 µm, VWR, USA), to get rid of cells and smaller particles present. Of the eluate, 300 µL was analysed using TRACE^{TM} 1310 Gas Chromatograph (Thermo Fisher Scientific, USA) with autosampler. Between every 5 samples, an external standard was measured, containing 0.2% formic acid and 1000 µM of the acids 2-methylvaleric acid, valeric acid, acetic acid, propionic acid, isobutyric acid and butyric acid. The external standard was used as a control to check that the peaks were not drifting and to observe possible variations between runs. Specifications about injector, column and detector, provided by PhD Morten Nilsen, is listed in appendix E.

2.8 Statistical analysis

Statistical analyses were conducted in RStudio version 1.3.1093 (RStudioTeam, 2020). Spearman correlation analyses were conducted to associate levels of bacteria present in the samples and levels of SCFAs. Analysis was performed with p-value less than 0.05 and the correlation was FDR corrected using the Holm's method (figure C.2).

3. Results

3.1 16S rRNA sequencing data

To identify and quantify the bacteria present in the faecal samples from the infants, the 16S rRNA genes of the cells present in the samples were sequenced using Illumina MiSeq. Hundred faecal samples from 6-month-old infants were sequenced. The data was analysed using QIIME pipeline, resulting in 283 OTUs. A sequencing depth of 5000 was set as threshold to retain sequences in the analysis, in accordance with rarefaction curve (figure F.1). Distribution of bacteria present in the infants is shown in figure 3.1, showing average values for the sample categories A (10 samples, partial analysis), B (five samples, complete analysis) and C (95 samples, no analysis).

Samples analysed were selected for high abundance of *R. gnavus*. The 15 samples chosen ranged from 19% to 59% abundance of *R. gnavus*, while the average amounts of *R. gnavus* in category A, B and C were 33%, 44% and 4%, respectively (figure G.1).

Sequencing showed great variation in abundance between the samples, both regarding abundance of *R. gnavus* and abundance of other bacteria present. The number of different bacteria with a relative abundance over 1% in the samples differed from six to 13 bacterial genus. As well as having a high relative abundance of *R. gnavus*, six of the samples were high in *Bifidobacterium* (>20%), three were high in *Bacteroides* (>20%), and three were high in *Clostridium* (>20%). Twelve samples were low (<1%) in one or two of the three, *Bifidobacterium, Bacteroides* and *Clostridium.* Abundance of *Bifidobacterium* ranged from 0% to 32%, abundance of *Bacteroides* ranged from 0% to 34%, and the abundance of *Clostridium* ranged from 0% to 35%.

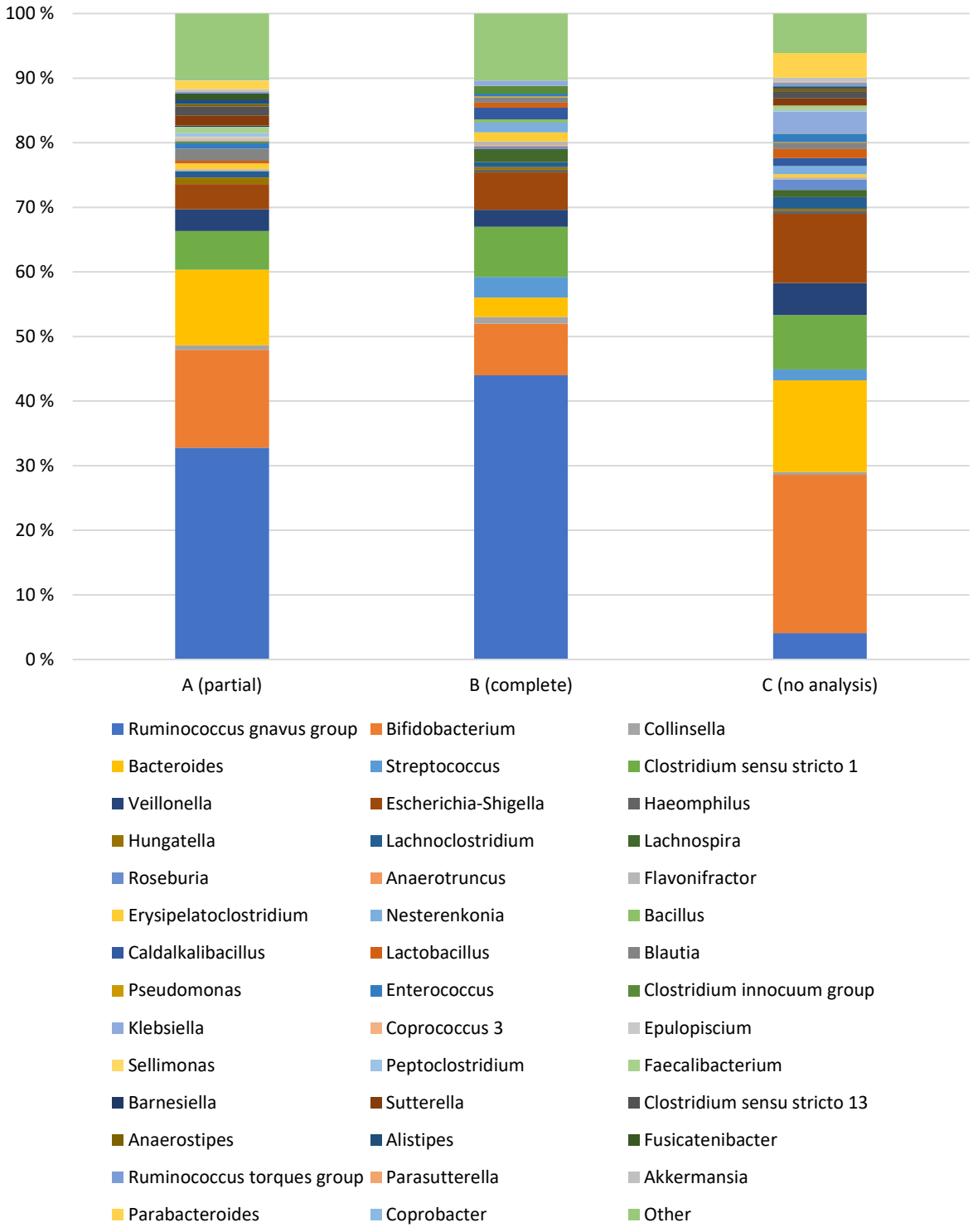

Figure 3.1. Distribution of bacteria in the infants. *Abundance of different bacterial taxa in 110 infants analysed, in percent based on 16S rRNA sequencing data. Infants 1-10 were grouped in category A and had been sequences previously by PhD Morten Nilsen, while infants 11-15 were grouped in category B, as only samples from infants in category B were collected without buffer (basic) and could be used for proteome analysis (complete analysis). The remaining 95 infants are grouped into category C, as their samples were not used for further analysis (no analysis). Average values of bacterial taxa abundance in the three categories are illustrated in the figure. Bacterial taxa with higher than 1% abundance in at least one infant are included in the figure. Less abundant taxa are included in "Other".*

3.2 Short chain fatty acid analysis

Levels of short chain fatty acids were measured in samples belonging to category A and B by gas chromatography. Samples of category C were not analysed using gas chromatography, as the samples were not used for shotgun sequencing or protein analysis. All samples were dominated by acetic acid, with an average of 76.1%, and all contained propionic acid (average 11.4%). The levels of SCFA differed between the samples, with one sample only containing acetic and propionic acid, and four samples only containing acetic, propionic and butyric acid (figure G.2). Figure 3.2 illustrates the average levels of SCFAs in the 15 samples measured.

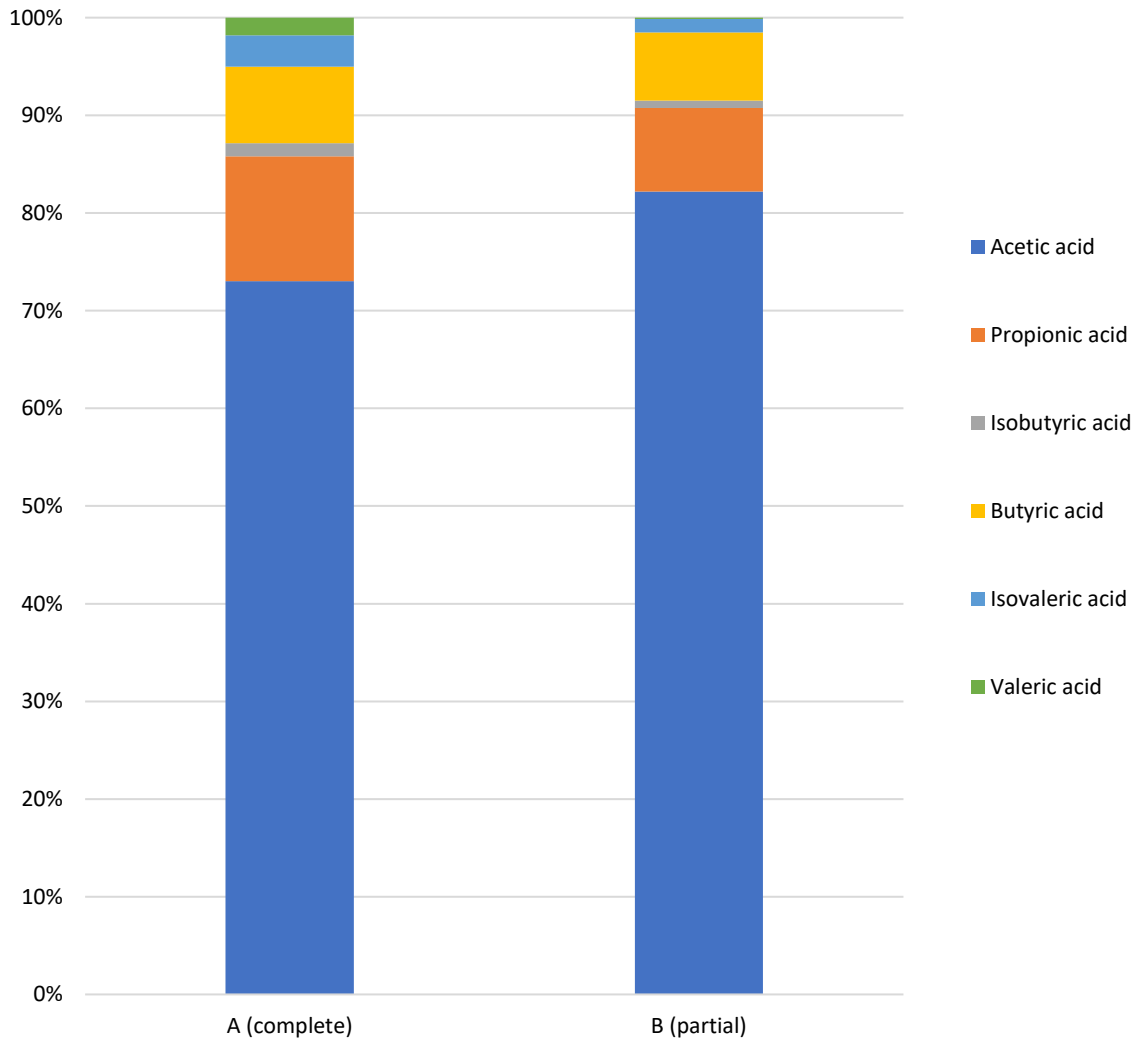

Figure 3.2. Distribution of short chain fatty acids. *Average levels of short chain fatty acids in the sample categories A and B in percent. Levels were measured by gas chromatography. Samples in category C were not analysed, as they were not to be used for any further analysis, neither shotgun sequencing, protein analysis or gene expression identification.*

Levels of propionic acid differed from 3.8% to 21.9% with an average level of 12.8% in category A and 8.5% in category B. Five samples were also high (>10%) in butyric acid, which ranged from 0.0% to 18.5% with an average level of 7.8% in category A and 7.0% in category B.

3.3 Correlation analysis of bacterial taxa and short chain fatty acids

The correlation between the SCFAs and the bacterial taxa were calculated using spearman correlation with FDR correction. Figure 3.3 is a correlogram showing correlation between bacteria and SCFAs present in the samples, with a significance level of 0.05 (table G.1).

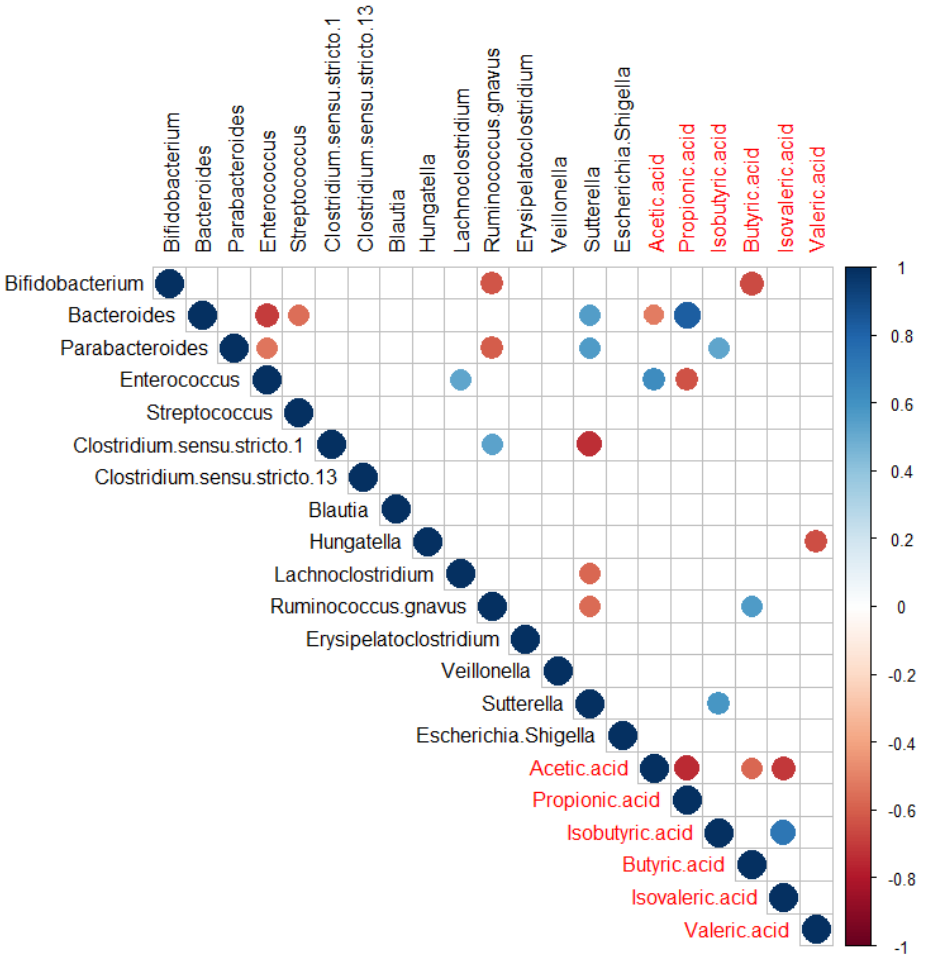

Figure 3.3. Correlation between SCFA and bacterial taxa. *The figure shows spearman correlation rho of only significant correlations (positive = blue, negative = red) between short chain fatty acids (red labels) and bacterial taxa (black labels), with α = 0.05. The correlations are FDR corrected. Size of circle reflects strength of correlation (rho), same as colour shading.*

Positive correlation was found between *R. gnavus* and *Clostridium* sensu stricto 1 (r = 0.539, p = 0.038) and butyric acid (r = 0.567, p = 0.027), while *R. gnavus* was negatively correlated with *Bifidobacterium* (r = -0.628, p = 0.012). Negative correlation was also found between *Bifidobacterium* and butyric acid (r = -0.653, p = 0.008). No correlation was found between *R. gnavus* and propionic acid, but there was a strong positive correlation between propionic acid and *Bacteroides* (r = 0.828, p = 0.0001).

3.4 Shotgun sequencing data

Samples from category A and B were shotgun sequenced, to access gene sequences present in the genome of *R. gnavus*. All samples were analysed together, and after filtration and annotation of sequences, 49 058 contigs remained, belonging to a total of 2110 different taxa, with 695 contigs being annotated to *R. gnavus*. All contigs belonging to *R. gnavus* were mapped to strain ATCC 29149. Amino acid sequences from the 695 contigs were gathered, resulting in 23 988 amino acid sequences (figure C.1). The amino acid sequences were used for further analysis of potential pathways for glycan degradation and SCFA production.

Annotation of proteins and identification of pathways based on the genome of *R. gnavus,* assembled through shotgun sequencing, was done using Ghost KOALA and KEGG. Since annotation and identification are based on the genome of *R. gnavus*, the results only illustrate the potential pathways and metabolic properties of *R. gnavus.* Using KEGG, 847 enzymes were annotated, and 13 complete pathways for carbohydrate metabolism were identified. Two complete pathways were found for SCFA production. Complete degradation pathways for three core molecules in host glycans were identified, in addition to 11 GHs involved in mucus and HMO degradation.

3.5 Proteomics

Proteins were isolated from the faecal samples of category B, and the proteins were identified using mass spectrometry. After protein isolation, the concentration of proteins in the samples was measured using BCA. The concentration in two samples from parallel 1 (11 and 14) and four samples from parallel 2 (11, 13, 14 and 15) were too low and could not be measured. The remaining samples in parallel 1 had the following concentrations: 0.285 µg/µL (12), 0.340 µg/µL (13) and 0.245 µg/µL (15). The sample which could be measured from parallel 2 had a

concentration of 0.010 µg/µL (12). All samples were concentrated as much as possible, but the protein concentration after concentration using SpeedVac was not measured and is not known. Protein concentration of samples added to the MS was measured using Nanodrop and is listed in table H.1. All samples were added to the SDS gel, and all samples were run on a mass spectrometer, independent of concentrations.

All contigs annotated as *R. gnavus* was used to annotate proteins using InterProScan. The protein sequences were thereafter used as a database. A total of 1921 protein sequences were identified from mass spectrometry. Perseus was used to filter away possible contaminants and human proteins, resulting in 959 protein sequences originating from gut microbes. Using Perseus, the two parallels of four out of five samples were clustered together according to hierarchical clustering, and the proteins were divided into two clusters according to presence of the proteins in the samples (figure H.1). The cluster containing the most abundant proteins contained 243 protein sequences, while the remaining 716 proteins belonged to the other cluster. Samples were normally distributed with respect to protein counts. There were some differences according to protein counts between parallel 1 and parallel 2, with generally more counts for parallel 1, but the distribution was approximately equal.

3.6 Proteins present in metabolic pathways

3.6.1 Short chain fatty acid production

Degradation of carbohydrates by bacteria in the gut of humans results in production of fermentation products, like short chain fatty acids. Using KEGG, pathways for both propionic acid and acetic acid production were identified in the *R. gnavus* genome. No genes or proteins involved in butyric acid production were found in the genome of *R. gnavus* or in the samples. The potential for propionic acid production was identified through the 1,2-propanediol pathway in KEGG (figure 3.4), while acetic acid production was found from acetyl-CoA, via acetyl phosphate.

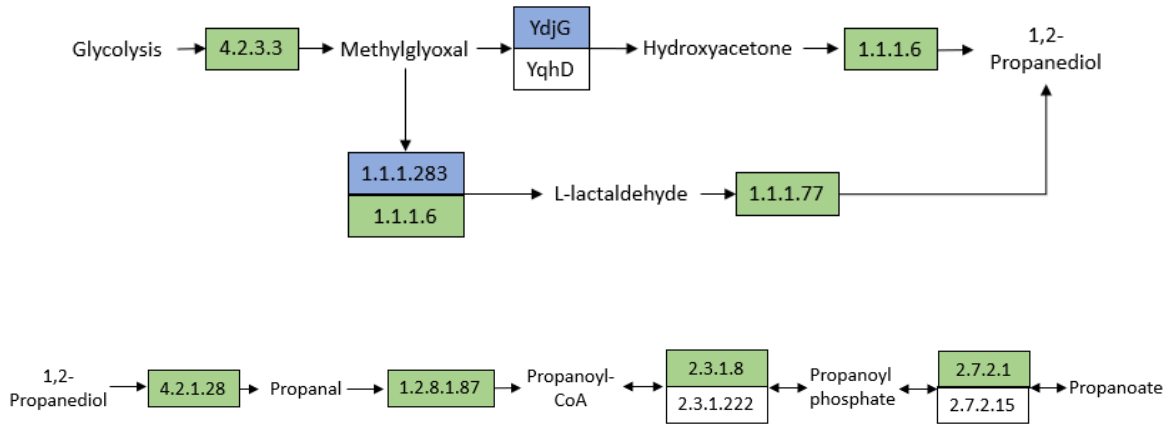

***Figure 3.4. Production of propionic acid**. Figure illustrates the 1,2-propanediol pathway for production of propionic acid from glycolysis products. Green proteins were present in both genome and proteome of R. gnavus, blue proteins were only present in the genome of R. gnavus, and not in the proteome, while white proteins were not present in either genome or proteome of R. gnavus. Illustration based on pathways obtained in KEGG.*

All four proteins necessary for propionic acid production from 1,2-propanediol were present in the proteome of two samples. Proteins necessary for production of 1,2-propanediol from glycolysis products were also present in the cells, all though two protein that could be used, but are not necessary for production, was missing from the proteome.

3.6.2 Host glycan degradation

When genes present in the genome of *R. gnavus* were annotated using KEGG, the genome of *R. gnavus* were found to code GH2 family β-galactosidases (EC:3.2.1.23), GH3 family β-glucosidases (EC:3.2.1.21), GH4 family α-galactosidases (EC:3.2.1.22), GH38 family α-mannosidases (EC:3.2.1.24), β-mannosidases (EC:3.2.1.25), GH101 family α-N-acetylgalactosaminidases and a GH20 family enzyme annotated as β-N-acetylhexosaminidase (EC:3.2.1.52), which all can break linkages in the core of host glycan molecules. Some of the enzymes are illustrated in figure 3.5. Of these proteins, GH2 family β-galactosidases were present in four of five *R. gnavus* proteomes. No GH20 family enzymes or β-mannosidases were found present in any *R. gnavus* proteome, while the other proteins listed were found in one to three of the samples. Degradation pathways were identified for glucose, galactose and GlcNAc, but not GalNAc.

Both human milk oligosaccharides and mucin molecules can be extended with fucose and/or sialic acid. Using KEGG annotations the genome of *R. gnavus* was found to harbour both GH29 and GH95 family fucosidases, and GH33 family sialidase, which can break off fucose and sialic acid, respectively, from the core structure of HMOs and mucin (figure 3.5).

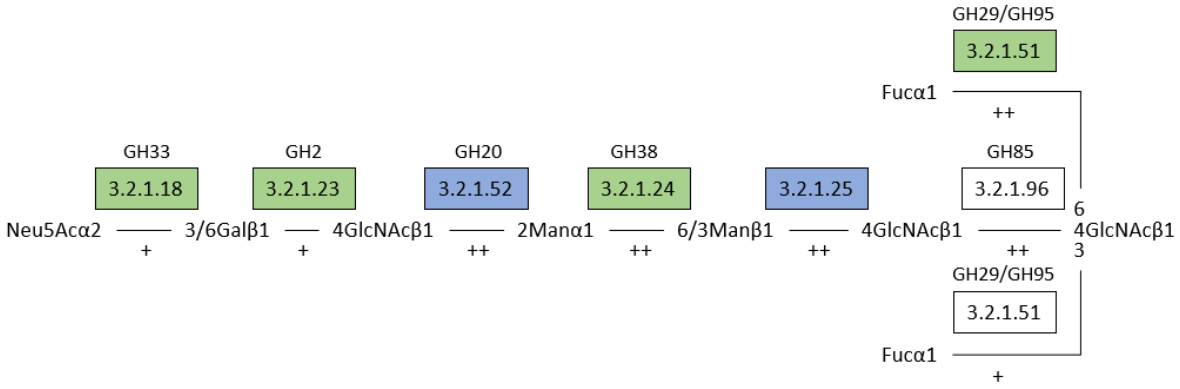

Figure 3.5. Glycan degradation. *The figure shows a N-glycan to illustrate the linkages between units in a glycan chain, and the EC-number and potential glycoside hydrolase family (GH) of the proteins necessary for breaking the bonds. Green EC-numbers were found present in the genome and proteome, blue EC-number were only found present in the genome, while white EC-numbers were not present, using KEGG. + = bond present in both HMO and mucin, ++ = bond only present in mucin.*

The gene *NEU1* (EC:3.2.1.18), coding sialidase-1, which can break off sialic acid bound to galactose or GalNAc by α-2,3 or α-2,6 linkages, was identified in the genome. In the *R. gnavus* proteome of two of five samples an extracellular GH33 family sialidase was identified. The protein sequence of the GH33 family sialidase matched that of an IT-sialidase previously found produced by *R. gnavus* (99.7% identity) (appendix I).

Genes and proteins involved in a pathway for sialic acid degradation were studied in *R. gnavus*. Five proteins are necessary to convert sialic acid to fructose-6-phosphate, coded by genes in the *nan*-cluster. Four out of five protein-coding genes were found in the genome of *R. gnavus*, and the proteins coded by these four genes were identified in the proteome, as illustrated in figure 3.6. The gene missing in the genome, *nanK*, codes a N-acylmannosamine kinase (EC:2.7.1.60), converting N-acetylmannosamine to N-acetylmannosamine 6-phosphate. There was also found a transport protein transporting extracellular N-acetylglucosamine into the cells (EC:2.7.1.193).

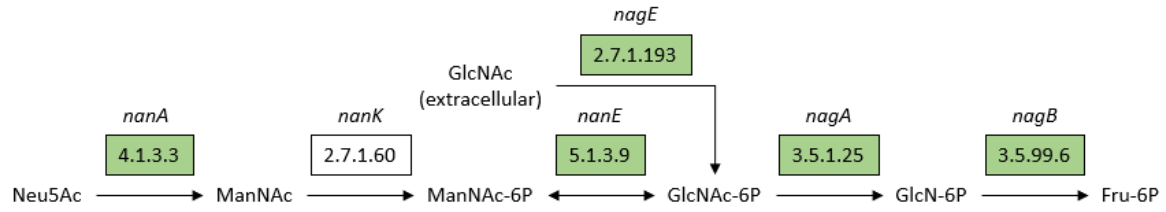

Figure 3.6. Sialic acid degradation. *Figure illustrates a pathway for sialic acid degradation, and the EC-number of proteins involved in the conversion of molecules. Green EC-numbers were found present in the genome and at least one proteome, while white EC-numbers were not present in the genome. Gene names are written above the EC-numbers.*

The proteome of one sample contained both a GH33 family sialidase and all proteins coded in the genome of *R. gnavus* for degradation of sialic acid (figure 3.6). The same sample also contained a transporter for sialic acid. Another sample contained the GH33 family sialidase, but none of the proteins necessary for degradation of sialic acid.

The genome of *R. gnavus* coded GH29 and GH95 family α-L-fucosidases which seemed to break the α-1,6-linkage between fucose and GlcNAc, according to KEGG (EC:3.2.1.51, figure 3.5). Several intracellular and extracellular fucosidases were identified in the proteome of all five samples (both parallels), where GH95 family fucosidases were present in all samples, while GH29 family fucosidases were present in only one sample. No fucose transporters were identified. Further, a degradation pathway from fucose to lactaldehyde and dihydroxyacetone phosphate was identified in the genome, where two of three genes were present, illustrated in figure 3.7. The gene product of these two genes, *FucI* and *FucA*, were present in the proteome of all five samples analysed. The gene coding L-fuculose kinase (*FucK*), converting L-fuculose into L-fuculose 1-phosphate, was lacking. No other pathways for fucose degradation were identified in the genome of *R. gnavus*.

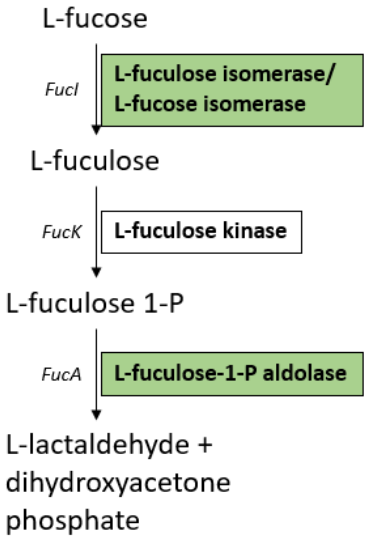

Figure 3.7. Fucose degradation. *Figure illustrates degradation of fucose, where genes for green proteins were found present in the genome and the proteins were found in all proteomes, while the gene for the white protein was not found in the genome and the protein was not identified in proteome. Gene names are written to the left.*

Lacto-N-tetraose (LNT) is a human milk oligosaccharide that can be degraded by the enzyme lacto-N-biosidase (EC:3.2.1.140, GH136), into lacto-N-biose (LNB) and lactose, illustrated in figure 3.8. The gene coding lacto-N-biosidase was not found present in the genome of *R. gnavus*. A gene coding 1,3-β-galactosyl-N-acetylhexosamine phosphorylase (lacto-N-biose phosphorylase, EC:2.4.1.211, GH112), breaking down LNB, was found present in the genome of *R. gnavus*, and the protein was present in the proteome of all five samples analysed. It was also found a gene coding β-galactosidase (EC:3.2.1.23, GH42), which can break down lactose. No membrane transport proteins for either LNB or lactose were identified in the genome.

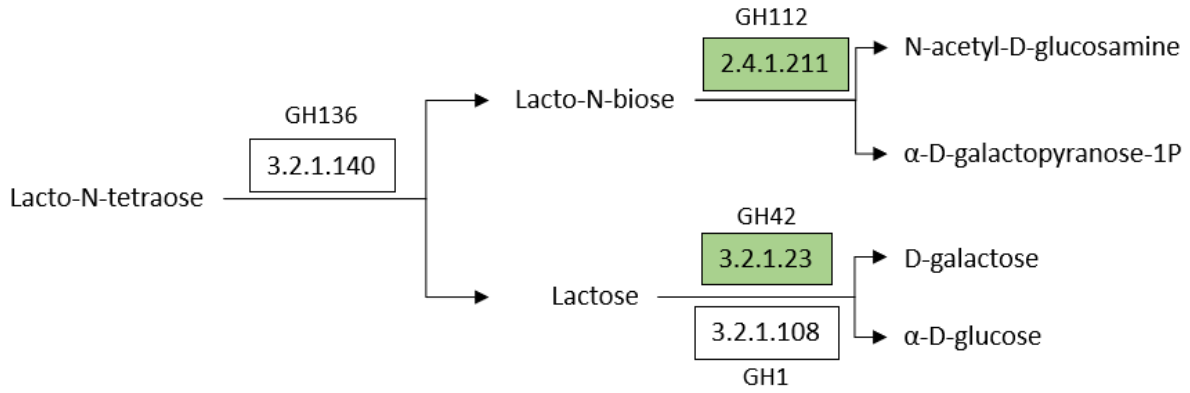

Figure 3.8. LNT degradation. *Illustration of lacto-N-tetraose degradation, into lacto-N-biose and lactose. The proteins necessary for degradations is illustrated with EC-numbers and GH-family number. Green EC-numbers were found present in the genome and proteome of R. gnavus, while white EC-numbers were not found in the genome or the proteome.*

3.7 Identification of gene expression

The expression of five genes, coding an α-L-fucosidase (RUMGNA_03833), α-galactosidase (RUMGNA_03611), β-galactosidase (RUMGNA_01638), N-acetylmannosamine-6-phosphate 2-epimerase (RUMGNA_02693) and a glyco_hyd_65N_2 domain-containing protein (RUMGNA_01058), were checked qualitatively in fourteen samples, using specific primers in qPCR. A difference of 2 cycles between samples in qPCR, with and without enzyme in cDNA synthesis, was set as a threshold for presence of gene in the samples. No genes were expressed in all samples. As shown in table 3.1, one sample had all five genes expressed.

Transcriptomics showed low expression of the 5 genes studied. Analysis showed low match between gene expression and proteome in the samples (table 3.1). No correlation was found between gene expression and the amount of *R. gnavus* or the SCFA levels in the samples. The expression of one gene (RUMGNA_03833) had a positive correlation with *Bifidobacterium* (r = 0.586, p = 0.027), but a negative correlation with *R. gnavus* (r = -0.574, p = 0.031).

Table 3.1. Identification of gene expression. *The table shows expression of five genes in 14 samples, and the percent amount of R. gnavus in the samples. The different colours illustrate the difference in cycles between samples with and without enzyme in cDNA synthesis. + = 2-4 cycles difference, ++ = 4-6 cycles difference, +++ = 6-10 cycles difference. The samples containing the associated proteins in the proteome are marked (†). Sample 1-10 belongs to category A, while sample 11-5 belong to category B.*

Sample	RUMGNA_02693 (Epimerase, *NanE*)	RUMGNA_03833 (α-L-fucosidase)	RUMGNA_01058 (Fucose activity)	RUMGNA_03611 (α-galactosidase)	RUMGNA_01638 (β-galactosidase)	% *R. gnavus*
1	+++	++	+	++	+	20
2		+				30
3		+++	+	++	+++	19
4				+		46
5	+	++		++	++	39
6		+		++	++	24
7						41
9						27
10						36
11				+		45
12		(†)				59
13	+	+				39
14	+ (†)				(†)	42
15	+	+++				35

4. Discussion

Few studies are available on the topic of HMO utilisation by *R. gnavus* in the infant gut, as previous studies have focused on the mucin degrading abilities of *R. gnavus* in adults. There are also few studies on how *R. gnavus* affects the SCFA composition of the infant and adult gut. A main finding of this thesis is that *R. gnavus* does not seem to be utilising entire HMO molecules alone, but mainly utilise mucin. Enzymes targeting mucin and not HMOs were identified in genome and proteome. Enzymes targeting both mucin and HMOs were also identified, indicating partial degradation of HMOs.

4.1 Potential mucin and human milk oligosaccharide utilisation by *R. gnavus*

4.1.1 Glycosyl hydrolases predict potential glycan degradation

β-galactosidases present in the proteome of *R. gnavus* can cleave galactose from glycans, but is not thought to cleave galactose from lactose. Extracellular β-galactosidase can break down lactose, a component of HMO consisting of glucose and galactose, which can be transported into the cell and utilised. All previously identified β-galactoses from *R. gnavus* has been predicted to be intracellular and *R. gnavus* has failed in growing solely on lactose in previous studies (Crost et al., 2013). As the cellular location of the β-galactosidases found in this thesis could not be determined using InterProScan and no transport proteins for lactose were identified in the genome using GhoastKOALA, the results together indicate that the β-galactosidases are intracellular and that *R. gnavus* cannot utilise lactose. The β-galactosidases seems to cleave galactose from glycan chains inside the cells.

The number of proteins with potential for mucin degradation in the genome and proteome of *R. gnavus* points towards degradation of mucin glycans by *R. gnavus*. GH3 family β-glucosidases, GH4 family α-galactosidases, GH38 family α-mannosidases and GH101 family α-N-acetylgalactosaminidases were present in some of the samples and will manly contribute to degradation of mucins and other glycoproteins (Hoskins et al., 1985; Zúñiga et al., 2018). Based on the findings of degradation pathways for glucose, galactose and GlcNAc, but not GalNAc, in the proteome of all samples, *R. gnavus* is thought to have the potential to utilise the three most abundant monosaccharides in the core of HMOs and two out of three most abundant monosaccharides in the core of mucins. In favour of these findings, the mucin

degrading enzymes together with the fucosidases, sialidases and β-galactosidases seem to contribute to mucin degradation by *R. gnavus*.

It does not seem that *R. gnavus* can utilise entire HMO molecules alone, but there might be some partial degradation and/or cross-feeding. This is supported by the absence of proteins belonging to GH20 family and other proteins associated with HMO degradation in *Bifidobacterium,* such as lacto-N-biosidase, N-acetylglucosaminidases and hexosaminidases, in the genome and proteome of *R. gnavus* (Sakanaka et al., 2020; Wada et al., 2008). One gene coding a GH20 family protein was found in the genome of *R. gnavus*, but the protein was not identified in the proteome of any sample. Several of the most important HMO degrading proteins in *Bifidobacterium* belongs to GH20 family (Sakanaka et al., 2020). The idea of exclusive degradation of HMO by *R. gnavus* is weakened by the absence of specific HMO degrading enzymes in the genome and proteome of the samples. Based on the findings of fucosidases, sialidases and β-galactosidases in the proteome, breaking linkages in both mucin and HMOs, *R. gnavus* might partially utilise HMOs and be involved in cross-feeding of different HMO derived carbohydrates.

4.1.2 Utilisation of sialic acid

R. gnavus in the infant gut is thought to release 2,7-anhydro-Neu5Ac instead of Neu5Ac from sialylated glycans, just like in the adult gut (Crost et al., 2013; Crost et al., 2016). This is based on the identification of a GH33 family exo-α-sialidase in the proteome of two samples, matching the amino acid sequence of an IT-sialidase identified in *R. gnavus* in previous studies (Crost et al., 2016; Tailford et al., 2015). The IT-sialidase has been shown to be specific for α-2,3-linkages in glycans and releases 2,7-anhydro-NeuAc, which is thought to give *R. gnavus* and advantage in mucus foraging. The presence of this type of sialidase could have been an explanation for the high abundance of *R. gnavus* in the samples. However, since sialidases were not present in all samples, this does not seem to be the case. Because of the low number of samples, it is not possible to predict if sialidase is expressed in most infants or not, and the importance of sialidase is difficult to anticipate.

Utilisation of sialic acid as carbon and/or energy source by *R. gnavus* does not have much support. *R. gnavus* was in previous studies found to have the complete *nan*-cluster necessary for sialic acid degradation present in the genome, but shotgun sequencing revealed the *nan*-

cluster to be non-complete in the infant samples studied in this thesis (Crost et al., 2013). The N-acylmannosamine kinase gene (*nanK*) was lacking in the genome, and translation of all *nan* genes only happened in one sample. Complete degradation of sialic acid in the cells does not have much support, and because of lack of evidence it is difficult to predict the importance of sialic acid degradation in *R. gnavus*. This is in contrast with the thought that utilisation of sialic acid is a reason for mucus adaptation in *R. gnavus* (Bell et al., 2019; Crost et al., 2016). The use and importance of sialic acid in infant and adults might be different, but an almost complete *nan*-cluster was identified and expressed in one sample, which indicates that the enzymes involved in sialic acid degradation have a function in the cells.

4.1.3 Utilisation of fucose

Fucosidases seems to be important for glycan utilisation in *R. gnavus*, and most fucosidases produced by *R. gnavus* seems to be favouring α-1,2-linkages to galactose, present in both HMO and mucin. Fucosidases identified in the proteome of *R. gnavus* were annotated with EC-number 3.2.1.51 by InterProScan and can be group into both GH29 and GH95 families. There are differences between the GH families, as GH29 mainly harbours α-1,3/1,4-L-fucosidases (EC:3.2.1.111) and GH95 mainly harbours α-1,2-L-fucosidases (EC:3.2.1.63)(Lombard et al., 2014). GH95 family fucosidases were present in all samples, while GH29 family fucosidase were only present in one sample. In favour of these observations fucosidases produced by *R. gnavus* has the potential to cleave fucose present on both HMO and mucin and seems essential for glycan utilisation, as they are present in all samples.

As with sialic acid degradation, evidence is lacking to determine if degradation of fucose by *R. gnavus* in the infant gut is happening. One essential gene in the degradation pathway of fucose, *FucK* coding L-fuculose kinase, was missing in the genome, while the other essential proteins were found in the proteome of all samples. If L-fuculose kinase had been present in the cells, fucose could be catabolised and used for propionic acid production. This has been shown in previous studies, where *R. gnavus* has utilised the HMOs 2-fucosyllactose and 3-fucosyllactose as sole carbon source and produced propionic acid (Crost et al., 2013). Fucose degradation is suggested to be more important than sialic acid degradation in the samples, as fucosidases and fucose degrading proteins were present in all proteomes of *R. gnavus* and since fucose is being suggested a significant reason for survival of *R. gnavus* on mucin alone

(Crost et al., 2013). However, further investigations are needed to identify complete degradation of fucose by *R. gnavus* in the infant gut.

4.1.4 Presence of lacto-N-biose phosphorylase in the genome of *R. gnavus*

The enzyme lacto-N-biose phosphorylase present in all proteomes of *R. gnavus* can contribute to glycan degradation by cleaving off galactose. Lacto-N-biose phosphorylase can break down LNB, an important building block of HMOs, releasing galactose and GlcNAc, but can also be used to cleave galactose from other glycans (Wada et al., 2008). Since LNT is broken down outside the cell and the lacto-N-biose phosphorylase has been shown to be situated inside the cells, LNB must be transported into the cell to be utilised (Wada et al., 2008). No transport proteins for LNB were found in the genome or the proteome. Based on the results it cannot be conclude if *R. gnavus* can utilise LNB originating from HMOs, or if lacto-N-biose phosphorylase is used to utilise galactose from mucins. Still, lacto-N-biose phosphorylase is believed to most likely cleave galactose from mucin derivates.

4.2 Short chain fatty acids in the infant gut

4.2.1 Positive correlation between *R. gnavus* and butyric acid

Positive correlation between the amount of *R. gnavus* and the levels of butyric acid in the samples can be explained by high abundance of butyric acid producing bacteria in the samples, such as *Clostridium* species. No pathways for butyric acid production were found in the genome of *R. gnavus*, pointing towards association with butyric acid producers. A positive correlation was found between *R. gnavus* and *Clostridium* sensu stricto 1, belonging to genus *Clostridium*, known for their butyric acid producing abilities (Van den Abbeele et al., 2013). This correlation can be part of the explanation for higher butyric acid levels. However, in a previous study *R. gnavus* was found to be negatively correlated with butyric acid, and the average level of butyric acid at six months were 4.1% (Nilsen et al., 2020). The sample material analysed were selected based on high levels of *R. gnavus* and might show less diversity in levels of SCFAs and bacteria than other study cohorts. This could be another explanation for the positive correlation between *R. gnavus* and butyric acid, and the results would in this case be uncertain. Because of the age of the infants studied, correlation with

butyric acid producers can be a reasonable explanation, as higher butyric acid levels are associated with a more mature gut (Nilsen et al., 2020).

4.2.2 Production of propionic acid by *R. gnavus*

Propionic acid seems to be produced by *R. gnavus* in the infant gut, although no correlation was found. This is based on the findings of a complete pathway for propionic acid production, the 1,2-propanediol pathway, in the genome and proteome of *R. gnavus*. Crost et al. (2013) has previously shown production of propionic acid when grown on fucosylated substrates and mucin, supporting production of propionic acid by *R. gnavus* in the infant gut. A previous study in our group measured average relative abundance of propionic acid in 100 samples from 6-month-old infants to be 6.8%, which is lower than measured in this experiment (11.4%)(Nilsen et al., 2020). As previously mentioned, the sample material analysed are selected and some diversity might be lost. High abundance of *R. gnavus* might still explain the high levels of propionic acid. Based on this observation there seems to be an association between high levels of *R. gnavus* and higher levels of propionic acid in infants, which is not illustrated in the results because of *R. gnavus* selected samples.

4.3 The potential role of *R. gnavus* in the infant gut

Bacteria belonging to *Bacteroides* has been shown to utilise HMOs by using mucin degrading enzymes, and the phenomenon might be used by *R. gnavus* in the gut (Marcobal et al., 2011). *Bacteroides* were, like *R. gnavus*, thought to be solely mucin utilising, but were shown by Marcobal et al. (2011) to upregulate the same genes when grown on mucin and HMO. The scientists introduced the idea that attracting bacteria utilising both mucin and HMO the first months of life might prepare the body for the transition into solid foods. This because many of the genes for mucin and HMO degrading enzymes in *Bacteroides* were localised together with genes coding GHs associated with degradation of plant polysaccharides. Many of the glycan degrading enzymes identified in the genome and proteome of *R. gnavus* can be used on both HMOs and mucins, and *R. gnavus* has previously been shown to grow solely on both mucin and HMO (Crost et al., 2013). It is difficult to state which of the two substrates are preferred by *R. gnavus* in the infant gut using the approach in this thesis. However, *R. gnavus* in the infant gut seems to partially utilise HMOs and use some of the HMO components, while mucin degradation might overall seem more important.

Based on the genome and proteome, *R. gnavus* is suggested to contribute to propionic acid production in the infant gut. All proteins necessary for propionic acid production through the 1,2-propanediol pathway were found present in the *R. gnavus* proteome. Propionic acid in the colon has been shown to have antimicrobial effects on pathogens, increase insulin sensitivity, and is thought to cause satiety, counteracting obesity (Al-Lahham et al., 2010). However, propionic acid has been shown to induce inflammations under certain environmental conditions. Propionic acid production by *R. gnavus* can contribute to stable glucose supply to the growing infant, as propionic acid can be absorbed by the host and converted to glucose through gluconeogenesis in the liver (Wong et al., 2006). Infants require more glucose, as brain development is very energy consuming, and propionic acid produced by *R. gnavus* can help fill this energy need.

R. gnavus was shown to be negatively correlated with *Bifidobacterium* when selected for high abundance of *R. gnavus*. This might indicate repression of favourable and "healthy" bacteria, like *Bifidobacteria*, by *R. gnavus* in the infant gut. Abundance of *R. gnavus* in the adult gut has been associated with disorders like Crohn's and eczema, and *R. gnavus* has been thought to induce inflammation in the gut (Hall et al., 2017; Henke et al., 2019; Zheng et al., 2016). No information has been found regarding if early colonisation of *R. gnavus* in the infant gut makes the infant more susceptible to *R. gnavus* associated diseases later in life, or if high abundance of *R. gnavus* early in life correlates with high abundance of *R. gnavus* in adulthood. What is known is that high abundance of *Bifidobacterium* in infants is beneficial and favourable. Repression of *Bifidobacteria* by *R. gnavus* could be caused by *R. gnavus* utilising HMOs, as HMOs are the primary energy and carbon source of *Bifidobacterium* species.

Early transitioning from an infant to an adult gut microbiota has previously been observed in infants born with caesarean section and has associated with development of allergies (Milani et al., 2017). Early transitioning has been characterised by a shift in levels of SCFA from dominance of acetic acid to higher levels of butyric and propionic acid (Nilsen et al., 2020). As *R. gnavus* seems to be associated with a more adult gut microbiota, this might indicate that the bacterium might have an unfavourable effect. A negative correlation was found between *R. gnavus* and *Bifidobacterium*, while a positive correlation was found with butyric acid. Both these observations are associated with a more adult and mature gut (Avershina et al., 2014; Nilsen et al., 2020). What is thought to be a normal, healthy infant gut, high in

Bifidobacterium and *Bacteroides*, starts to move towards an adult microbiota around age one to two, around the time of weaning (Nilsen et al., 2020; Radjabzadeh et al., 2020). A microbiota resembling an adult microbiota at 6 months is therefore not desirable.

Although high abundance of *R. gnavus* is thought to have a negative impact on the gut health in infants, there might be some positive aspect. High abundance of *R. gnavus* seems to be associated with higher levels of propionic acid in the gut. Propionic acid can be absorbed by the epithelium and transported to the liver, where it is used in gluconeogenesis. *R. gnavus* might therefore contribute to increased glucose levels in the liver, which might be beneficial under certain conditions. Under conditions where the glucose supply is limiting, high abundance of *R. gnavus* producing propionic acid can secure regular supply of glucose to the growing infant brain. Limited supply of glucose is usually not a problem in the western world, and therefore the negative aspects of harbouring *R. gnavus* might have more impact on the infant's health and development. High abundance of *R. gnavus* is therefore, overall, thought to be unfavourable.

4.4 Technical discussion

4.4.1 Protein isolation and analysis

The protein analysis was to be done in two technical replicates, but was not due to problems with equipment. Two different filtering approaches were used to isolate bacterial cells, and the two approaches affected the results slightly. There were some differences in proteins identified in the parallels and in the protein counts. However, the parallels were clustered together, strengthening the results and interpretations (figure H.1). Parallels are used in experiments to verify results. The presence of a protein in both parallels can be trusted more than if the protein is just present in one parallel. In the case of this analysis the parallels were not technical replicates, and it can be assumed that more bacterial cells were lost in the processing of parallel 2 as bacterial cells were isolated by centrifugation and not filtering. Some differences in presence of proteins must therefore be expected. In the analysis a protein is assumed present in the cells if it was detected in the MS/MS run. Both proteins present in both parallels and proteins present in just one parallel are included in the analysis. Almost all proteins discussed were present in more than one sample, and all proteins mentioned were coded in the genome of *R. gnavus*. There is a chance that proteins identified in reality is not

present in the cells, but this must be investigated in further studies by for instance growth experiments.

The protein extraction protocol used will only isolate proteins present inside the cells at the sampling time, and not proteins elsewhere in the samples. Extracellular proteins secreted from the bacterial cells will not be detected, but extracellular proteins not yet secreted can be detected. Transport proteins anchored to the cell membrane and cell wall is generally difficult to isolate, as they are tightly bound, and might therefore be excluded in the isolation process. This means that some protein potentially involved in HMO and/or mucin degradation and transport will be lost. The concentration of extracted proteins was low, but because of high concentrations of salt in the lysis buffer the samples could not be concentrated sufficiently for optimised results on mass spectrometer. Lower concentrations of proteins can have affected the run of the mass spectrometer, and therefore the results.

Annotation of protein sequences identified from mass spectrometry was done using a self-constructed database. This method will only annotate protein sequences coded in the genome of *R. gnavus*. Proteins produced by other bacteria present in the samples could be matched with gene sequences in *R. gnavus*, based on similar protein sequences. This approach was chosen to simplify annotation of proteins and to easier identify proteins belonging to *R. gnavus* in the metaproteome of the samples. Extracting proteins belonging to *R. gnavus* based on the metagenome of the samples would be very time consuming. The approach will therefore result in higher risk of false positive results.

4.4.2 Measuring short chain fatty acid levels

The absolute concentrations of SCFA in the samples could not be measured, as weight of the samples varied. The dilutions of the samples were therefore not known, and only relative levels of SCFA could be measured. When working with gas chromatographs there can be difficulties, and there has been problems with the instrument in the lab at previous occasions. There is a chance of false peaks and that the peaks move around on the chromatogram. External and internal controls were used to monitor the location of the peaks throughout the run. The peaks were believed to be correct, as no shifting in position was observed for the internal control, and the external control gave even results throughout the run.

4.4.3 Analysis of RNA

The amount of RNA present in the samples was below the detection limit. Generally low concentrations of RNA could affect the qPCR results, and gene expression data must be considered uncertain. Because of high C_q-values, three categories were made to easier analyse the gene expression data (table 3.1). There also seemed to be a cross reaction with the primers, as one primer was positively correlated with *Bifidobacterium*, and no connection was found between the gene expression and *R. gnavus*. Low concentration of RNA and not 100% specificity for *R. gnavus* makes the results unreliable, and it would be necessary to do a full transcriptomics analysis with RNA sequencing to determine the association between gene transcription and protein translation.

5. Conclusion and further research

Results obtained from this thesis suggests that the gut microbe *R. gnavus* has the potential to utilise both mucin glycans and HMOs as energy source and seems to contribute to propionic acid production in the infant gut. *R. gnavus* seems to utilise more mucins than HMOs in the infant gut, based on the absence of specific enzymes for HMO utilisation in genome and proteome and presence of mucin specific enzymes. Evidence for complete degradation of sialic acid and fucose is lacking, as some genes were missing in the pathways. If sialic acid and fucose cannot be broken down by *R. gnavus* in the infant gut, they might be cleaved off by sialidases and fucosidases coded in the *R. gnavus* genome to access core carbohydrates of mucin glycans and/or HMOs. High abundance of *R. gnavus* in the infant gut may have both positive and negative effects, but in conclusion, high abundance of *R. gnavus* in the infant gut is considered to be unfavourable.

Several knowledge gaps still need to be filled concerning preferred substrate and degradation of sialic acid and fucose by *R. gnavus*. Proteins with potential to partially degrade both mucin glycans and HMOs have been found, making it difficult to determine preferred substrate of *R. gnavus* in the infant gut. Cultivation experiments should be conducted to understand which substrates *R. gnavus* utilises, where isolated *R. gnavus* cells from infant faecal samples are grown on different types of HMOs and mucins. A more extensive transcriptomics analysis should also be conducted, using RNA sequencing. Controls and contrasts should be included to identify what is special about high abundance of *R. gnavus*, and what the effect of *R. gnavus* on the gut is. Interesting aspects to consider, and where knowledge is missing, would also be the effect of diet on *R. gnavus* metabolism, requiring more information about amount of breastfeeding and weaning than available for these samples. The effect of high abundance of *R. gnavus* in infants on the health and quality of life in adulthood is also interesting, as *R. gnavus* has been associated with disorders, including dysbiosis, in the adult gut. Filling the knowledge gaps considering metabolism and action of *R. gnavus* will reveal the importance of *R. gnavus* in the gut. This will lead us one step closer to confirm if high abundance of *R. gnavus* in the infant gut is beneficial or harmful.

6. References

Aagaard, K., Ma, J., Antony, K. M., Ganu, R., Petrosino, J. & Versalovic, J. (2014). The placenta harbors a unique microbiome. *Science translational medicine*, 6 (237): 237ra65-237ra65. doi: 10.1126/scitranslmed.3008599.

Aebersold, R. & Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature*, 537 (7620): 347-355. doi: 10.1038/nature19949.

Al-Lahham, S. a. H., Peppelenbosch, M. P., Roelofsen, H., Vonk, R. J. & Venema, K. (2010). Biological effects of propionic acid in humans; metabolism, potential applications and underlying mechanisms. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, 1801 (11): 1175-1183. doi: https://doi.org/10.1016/j.bbalip.2010.07.007.

Arntzen, M. Ø., Karlskås, I. L., Skaugen, M., Eijsink, V. G. H. & Mathiesen, G. (2015). Proteomic Investigation of the Response of *Enterococcus faecalis* V583 when Cultivated in Urine. *PLoS ONE*, 10 (4). doi: https://doi.org/10.1371/journal.pone.0126694.

Avershina, E., Storrø, O., Øien, T., Johnsen, R., Pope, P. & Rudi, K. (2014). Major faecal microbiota shifts in composition and diversity with age in a geographically restricted cohort of mothers and their children. *FEMS Microbiology Ecology*, 87 (1): 280-290. doi: 10.1111/1574-6941.12223.

Avershina, E., Lundgård, K., Sekelja, M., Dotterud, C., Storrø, O., Øien, T., Johnsen, R. & Rudi, K. (2016). Transition from infant- to adult-like gut microbiota. *Environmental Microbiology*, 18 (7): 2226-2236. doi: https://doi.org/10.1111/1462-2920.13248.

Barcenilla, A., Pryde, S. E., Martin, J. C., Duncan, S. H., Stewart, C. S., Henderson, C. & Flint, H. J. (2000). Phylogenetic Relationships of Butyrate-Producing Bacteria from the Human Gut. *Applied and Environmental Microbiology*, 66 (4): 1654-1661. doi: 10.1128/aem.66.4.1654-1661.2000.

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

Besten, G. d., Lange, K., Havinga, R., Dijk, T. H. v., Gerding, A., Eunen, K. v., Müller, M., Groen, A. K., Hooiveld, G. J., Bakker, B. M., et al. (2013). Gut-derived short-chain fatty acids are vividly assimilated into host carbohydrates and lipids. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 305 (12): G900-G910. doi: 10.1152/ajpgi.00265.2013.

Blank, D., Dotz, V., Geyer, R. & Kunz, C. (2012). Human milk oligosaccharides and Lewis blood group: individual high-throughput sample profiling to enhance conclusions from functional studies. *Advances in nutrition (Bethesda, Md.)*, 3 (3): 440S-9S. doi: 10.3945/an.111.001446.

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

Bolger, A. M., Lohse, M. & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, 30 (15): 2114-2120. doi: 10.1093/bioinformatics/btu170.

Boom, R., Sol, C. J., Salimans, M. M., Jansen, C. L., Wertheim-van Dillen, P. M. & van der Noordaa, J. (1990). Rapid and simple method for purification of nucleic acids. *Journal of clinical microbiology*, 28 (3): 495-503. doi: 10.1128/JCM.28.3.495-503.1990.

Bullich, C., Keshavarzian, A., Garssen, J., Kraneveld, A. & Perez-Pardo, P. (2019). Gut Vibes in Parkinson's Disease: The Microbiota-Gut-Brain Axis. *Movement Disorders Clinical Practice*, 6 (8): 639-651. doi: https://doi.org/10.1002/mdc3.12840.

Casén, C., Vebø, H. C., Sekelja, M., Hegge, F. T., Karlsson, M. K., Ciemniejewska, E., Dzankovic, S., Frøyland, C., Nestestog, R., Engstrand, L., et al. (2015). Deviations in human gut microbiota: a novel diagnostic test for determining dysbiosis in patients with IBS or IBD. *Alimentary Pharmacology & Therapeutics*, 42 (1): 71-83. doi: https://doi.org/10.1111/apt.13236.

Cheng, J., Ringel-Kulka, T., Heikamp-de Jong, I., Ringel, Y., Carroll, I., de Vos, W. M., Salojärvi, J. & Satokari, R. (2016). Discordant temporal development of bacterial phyla and the emergence of core in the fecal microbiota of young children. *The ISME Journal*, 10 (4): 1002-1014. doi: 10.1038/ismej.2015.177.

Cox, J. & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26 (12): 1367-1372. doi: 10.1038/nbt.1511.

Cravatt, B. F., Simon, G. M. & Yates Iii, J. R. (2007). The biological impact of mass-spectrometry-based proteomics. *Nature*, 450 (7172): 991-1000. doi: 10.1038/nature06525.

Crost, E. H., Tailford, L. E., Le Gall, G., Fons, M., Henrissat, B. & Juge, N. (2013). Utilisation of Mucin Glycans by the Human Gut Symbiont Ruminococcus gnavus Is Strain-Dependent. *PLOS ONE*, 8 (10): e76341. doi: 10.1371/journal.pone.0076341.

Crost, E. H., Tailford, L. E., Monestier, M., Swarbreck, D., Henrissat, B., Crossman, L. C. & Juge, N. (2016). The mucin-degradation strategy of Ruminococcus gnavus: The importance of intramolecular trans-sialidases. *Gut Microbes*, 7 (4): 302-312. doi: 10.1080/19490976.2016.1186334.

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): 11971-11975. doi: 10.1073/pnas.1002601107.

Donohoe, D. R., Garge, N., Zhang, X., Sun, W., O'Connell, T. M., Bunger, M. K. & Bultman, S. J. (2011). The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon. *Cell metabolism*, 13 (5): 517-526. doi: 10.1016/j.cmet.2011.02.018.

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S. R., Nelson, K. E. & Relman, D. A. (2005). Diversity of the Human Intestinal Microbial Flora. *Science*, 308 (5728): 1635-1638. doi: 10.1126/science.1110591.

Eiwegger, T., Stahl, B., Schmitt, J., Boehm, G., Gerstmayr, M., Pichler, J., Dehlink, E., Loibichler, C., Urbanek, R. & Szépfalusi, Z. (2004). Human Milk–Derived Oligosaccharides and Plant-Derived Oligosaccharides Stimulate Cytokine Production of Cord Blood T-Cells In Vitro. *Pediatric Research*, 56 (4): 536-540. doi: 10.1203/01.PDR.0000139411.35619.B4.

Engfer, M. B., Stahl, B., Finke, B., Sawatzki, G. & Daniel, H. (2000). Human milk oligosaccharides are resistant to enzymatic hydrolysis in the upper gastrointestinal tract. *The American Journal of Clinical Nutrition*, 71 (6): 1589-1596. doi: 10.1093/ajcn/71.6.1589.

Faith, J. J., Guruge, J. L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A. L., Clemente, J. C., Knight, R., Heath, A. C., Leibel, R. L., et al. (2013). The long-term stability of the human gut microbiota. *Science*, 341 (6141): 1237439. doi: 10.1126/science.1237439.

Favier, C. F., Vaughan, E. E., De Vos, W. M. & Akkermans, A. D. L. (2002). Molecular Monitoring of Succession of Bacterial Communities in Human Neonates. *Applied and Environmental Microbiology*, 68 (1): 219-226. doi: 10.1128/aem.68.1.219-226.2002.

Ganapathy, V., Thangaraju, M., Prasad, P. D., Martin, P. M. & Singh, N. (2013). Transporters and receptors for short-chain fatty acids as the molecular link between colonic bacteria and the host. *Current Opinion in Pharmacology*, 13 (6): 869-874. doi: https://doi.org/10.1016/j.coph.2013.08.006.

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.

Geuking, M. B., Köller, Y., Rupp, S. & McCoy, K. D. (2014). The interplay between the gut microbiota and the immune system. *Gut Microbes*, 5 (3): 411-418. doi: 10.4161/gmic.29330.

Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. (1999). Correlation between Protein and mRNA Abundance in Yeast. *Molecular and Cellular Biology*, 19 (3): 1720-1730. doi: 10.1128/mcb.19.3.1720.

Hall, A. B., Yassour, M., Sauk, J., Garner, A., Jiang, X., Arthur, T., Lagoudas, G. K., Vatanen, T., Fornelos, N., Wilson, R., et al. (2017). A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Medicine*, 9 (1): 103. doi: 10.1186/s13073-017-0490-5.

Henke, M. T., Kenny, D. J., Cassilly, C. D., Vlamakis, H., Xavier, R. J. & Clardy, J. (2019). *Ruminococcus gnavus*, a member of the human gut microbiome associated with Crohn's disease, produces an inflammatory polysaccharide. *Proceedings of the National Academy of Sciences*, 116 (26): 12672-12677. doi: 10.1073/pnas.1904099116.

Hoskins, L. C., Agustines, M., McKee, W. B., Boulding, E. T., Kriaris, M. & Niedermeyer, G. (1985). Mucin degradation in human colon ecosystems. Isolation and properties of fecal strains that degrade ABH blood group antigens and oligosaccharides from mucin glycoproteins. *J Clin Invest*, 75 (3): 944-53. doi: 10.1172/jci111795.

Huang, H. (2014). *QIIME Workflow*. Available at: https://sites.google.com/site/knightslabwiki/qiime-workflow (accessed: 17.03.21).

Hughes, H. K., Rose, D. & Ashwood, P. (2018). The Gut Microbiota and Dysbiosis in Autism Spectrum Disorders. *Current Neurology and Neuroscience Reports*, 18 (11): 81. doi: 10.1007/s11910-018-0887-6.

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

Jiang, H.-y., Zhang, X., Yu, Z.-h., Zhang, Z., Deng, M., Zhao, J.-h. & Ruan, B. (2018). Altered gut microbiota profile in patients with generalized anxiety disorder. *Journal of Psychiatric Research*, 104: 130-136. doi: https://doi.org/10.1016/j.jpsychires.2018.07.007.

Jiménez, E., Fernández, L., Marín, M. L., Martín, R., Odriozola, J. M., Nueno-Palop, C., Narbad, A., Olivares, M., Xaus, J. & Rodríguez, J. M. (2005). Isolation of Commensal Bacteria from Umbilical Cord Blood of Healthy Neonates Born by Cesarean Section. *Current Microbiology*, 51 (4): 270-274. doi: 10.1007/s00284-005-0020-3.

Johansson, M. E. V., Ambort, D., Pelaseyed, T., Schütte, A., Gustafsson, J. K., Ermund, A., Subramani, D. B., Holmén-Larsson, J. M., Thomsson, K. A., Bergström, J. H., et al. (2011). Composition and functional role of the mucus layers in the intestine. *Cellular and Molecular Life Sciences*, 68 (22): 3635. doi: 10.1007/s00018-011-0822-3.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30 (9): 1236-1240. doi: 10.1093/bioinformatics/btu031.

Joossens, M., Huys, G., Cnockaert, M., De Preter, V., Verbeke, K., Rutgeerts, P., Vandamme, P. & Vermeire, S. (2011). Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. *Gut*, 60 (5): 631-637. doi: 10.1136/gut.2010.223263.

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

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H. & Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7: e7359-e7359. doi: 10.7717/peerj.7359.

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

Lauder, A. P., Roche, A. M., Sherrill-Mix, S., Bailey, A., Laughlin, A. L., Bittinger, K., Leite, R., Elovitz, M. A., Parry, S. & Bushman, F. D. (2016). Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. *Microbiome*, 4 (1): 29. doi: 10.1186/s40168-016-0172-3.

Lesk, A. M. (2016). *Introduction to Protein Science*. Third ed. UK: Oxford University Press.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25 (16): 2078-2079. doi: 10.1093/bioinformatics/btp352.

Linde.AG. (2021a). *Flame ionisation detector.* Pullach, Germany: Linde Gases Division. Available at: http://hiq.linde-gas.com/en/images/Application%20note_Flame%20Ionisation%20Detector_tcm899-92468.pdf (accessed: 10.02.21).

Linde.AG. (2021b). *Gas Chromatography*. Pullach, Germany: Linde Gases Division. Available at: http://hiq.linde-gas.com/en/images/Application%20sheetHiQ_Gas_Chromatography%28appl%29_tcm899-92473.pdf (accessed: 10.02.21).

Lombard, V., Ramulu, H. G., Drula, E., Coutinho, P. & Henrissat, B. (2014). The Carbohydrate-active enzymes database (CAZy) (no. 10.1093/nar/gkt1178). Available at: http://www.cazy.org/.

Louis, P. & Flint, H. J. (2017). Formation of propionate and butyrate by the human colonic microbiota. *Environmental Microbiology*, 19 (1): 29-41. doi: 10.1111/1462-2920.13589.

Marcobal, A., Barboza, M., Sonnenburg, E. D., Pudlo, N., Martens, E. C., Desai, P., Lebrilla, C. B., Weimer, B. C., Mills, D. A., German, J. B., et al. (2011). Bacteroides in the infant gut consume milk oligosaccharides via mucus-utilization pathways. *Cell host & microbe*, 10 (5): 507-514. doi: 10.1016/j.chom.2011.10.007.

Marcobal, A. & Sonnenburg, J. L. (2012). Human milk oligosaccharide consumption by intestinal microbiota. *Clinical Microbiology and Infection*, 18: 12-15. doi: https://doi.org/10.1111/j.1469-0691.2012.03863.x.

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. *Microbiology and Molecular Biology Reviews*, 81 (4): e00036-17. doi: 10.1128/mmbr.00036-17.

Moore, W. E. C., Johnson, J. L. & Holdeman, L. V. (1976). Emendation of Bacteroidaceae and Butyrivibrio and Descriptions of Desulfomonas gen. nov. and Ten New Species in the Genera Desulfomonas, Butyrivibrio, Eubacterium, Clostridium, and Ruminococcus. *International Journal of Systematic and Evolutionary Microbiology*, 26 (2): 238-252. doi: https://doi.org/10.1099/00207713-26-2-238.

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.

Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome research*, 27 (5): 824-834. doi: 10.1101/gr.213959.116.

Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME Journal*, 11 (12): 2864-2868. doi: 10.1038/ismej.2017.126.

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

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490 (7418): 55-60. doi: 10.1038/nature11450.

Radjabzadeh, D., Boer, C. G., Beth, S. A., van der Wal, P., Kiefte-De Jong, J. C., Jansen, M. A. E., Konstantinov, S. R., Peppelenbosch, M. P., Hays, J. P., Jaddoe, V. W. V., et al. (2020). Diversity, compositional and functional differences between gut microbiota of children and adults. *Scientific Reports*, 10 (1): 1040. doi: 10.1038/s41598-020-57734-z.

Ringel-Kulka, T., Cheng, J., Ringel, Y., Salojärvi, J., Carroll, I., Palva, A., de Vos, W. M. & Satokari, R. (2013). Intestinal microbiota in healthy U.S. young children and adults--a high throughput microarray analysis. *PloS one*, 8 (5): e64315-e64315. doi: 10.1371/journal.pone.0064315.

RStudioTeam. (2020). *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, PBC. Available at: http://www.rstudio.com/.

Sagheddu, V., Patrone, V., Miragoli, F., Puglisi, E. & Morelli, L. (2016). Infant Early Gut Colonization by Lachnospiraceae: High Frequency of Ruminococcus gnavus. *Frontiers in Pediatrics*, 4 (57). doi: 10.3389/fped.2016.00057.

Sakanaka, M., Gotoh, A., Yoshida, K., Odamaki, T., Koguchi, H., Xiao, J.-z., Kitaoka, M. & Katayama, T. (2020). Varied Pathways of Infant Gut-Associated Bifidobacterium to Assimilate Human Milk Oligosaccharides: Prevalence of the Gene Set and Its Correlation with Bifidobacteria-Rich Microbiota Formation. *Nutrients*, 12 (1): 71.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74 (12): 5463-5467. doi: 10.1073/pnas.74.12.5463.

Schadt, E. E., Turner, S. & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, 19 (R2): R227-R240. doi: 10.1093/hmg/ddq416.

Tailford, L. E., Owen, C. D., Walshaw, J., Crost, E. H., Hardy-Goddard, J., Le Gall, G., de Vos, W. M., Taylor, G. L. & Juge, N. (2015). Discovery of intramolecular trans-sialidases in human gut microbiota suggests novel mechanisms of mucosal adaptation. *Nature communications*, 6: 7624-7624. doi: 10.1038/ncomms8624.

Tsukuda, N., Yahagi, K., Hara, T., Watanabe, Y., Matsumoto, H., Mori, H., Higashi, K., Tsuji, H., Matsumoto, S., Kurokawa, K., et al. (2021). Key bacterial taxa and metabolic pathways affecting gut short-chain fatty acid profiles in early life. *The ISME Journal*. doi: 10.1038/s41396-021-00937-7.

Urashima, T., Saito, T., Nakamura, T. & Messer, M. (2001). Oligosaccharides of milk and colostrum in non-human mammals. *Glycoconjugate Journal*, 18 (5): 357-371. doi: 10.1023/A:1014881913541.

Van den Abbeele, P., Belzer, C., Goossens, M., Kleerebezem, M., De Vos, W. M., Thas, O., De Weirdt, R., Kerckhof, F.-M. & Van de Wiele, T. (2013). Butyrate-producing Clostridium cluster XIVa species specifically colonize mucins in an in vitro gut model. *The ISME Journal*, 7 (5): 949-961. doi: 10.1038/ismej.2012.158.

van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, 30 (9): 418-426. doi: https://doi.org/10.1016/j.tig.2014.07.001.

Vargas-Albores, F., Ortiz-Suárez, L. E., Villalpando-Canchola, E. & Martínez-Porchas, M. (2017). Size-variable zone in V3 region of 16S rRNA. *RNA biology*, 14 (11): 1514-1521. doi: 10.1080/15476286.2017.1317912.

Wada, J., Ando, T., Kiyohara, M., Ashida, H., Kitaoka, M., Yamaguchi, M., Kumagai, H., Katayama, T. & Yamamoto, K. (2008). Bifidobacterium bifidum Lacto-N-Biosidase, a Critical Enzyme for the Degradation of Human Milk Oligosaccharides with a Type 1 Structure. *Applied and Environmental Microbiology*, 74 (13): 3996-4004. doi: 10.1128/aem.00149-08.

Wang, B., McVeagh, P., Petocz, P. & Brand-Miller, J. (2003). Brain ganglioside and glycoprotein sialic acid in breastfed compared with formula-fed infants. *The American Journal of Clinical Nutrition*, 78 (5): 1024-1029. doi: 10.1093/ajcn/78.5.1024.

Whitman, W. B., Coleman, D. C. & Wiebe, W. J. (1998). Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences*, 95 (12): 6578-6583. doi: 10.1073/pnas.95.12.6578.

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): 5088-5090. doi: 10.1073/pnas.74.11.5088.

Wong, J. M. W., de Souza, R., Kendall, C. W. C., Emam, A. & Jenkins, D. J. A. (2006). Colonic Health: Fermentation and Short Chain Fatty Acids. *Journal of Clinical Gastroenterology*, 40 (3): 235-243.

Wood, D. E. & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15 (3): R46. doi: 10.1186/gb-2014-15-3-r46.

Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A. & Singer, S. W. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, 2 (1): 26. doi: 10.1186/2049-2618-2-26.

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.

Zheng, H., Liang, H., Wang, Y., Miao, M., Shi, T., Yang, F., Liu, E., Yuan, W., Ji, Z. & Li, D. (2016). Altered Gut Microbiota Composition Associated with Eczema in Infants. *PLoS ONE*, 11 (11). doi: https://doi.org/10.1371/journal.pone.0166026.

Zúñiga, M., Monedero, V. & Yebra, M. J. (2018). Utilization of Host-Derived Glycans by Intestinal Lactobacillus and Bifidobacterium Species. *Frontiers in microbiology*, 9: 1917-1917. doi: 10.3389/fmicb.2018.01917.

Appendix

Appendix A: Experimental setup

Sample info

The initial plan of the thesis was to study the genome, transcriptome, proteome and SCFA levels of 10 samples. These 10 samples had not been collected without buffer, making it impossible to do proteomics on the samples. Therefore, we had to find five new samples, collected in both DNA stabilising buffer, DNA/RNA shield buffer and without buffer. This complicated the setup of the experiment, as a total of 15 samples were analysed. The genome, transcriptome, proteome and SCFA levels could only be compared in five samples, while the 10 remaining samples could be compared according to genome, SCFA levels and partly transcriptome. The ideal set up was to analyse the proteome of all samples, as this gives information on what is happening in the samples. Samples analysed were all high in *R. gnavus* and we did not include any contrasts. We could therefore not compare findings from samples high in *R. gnavus* with samples with less *R. gnavus*, which, in hindsight, we probably should have included.

RNA primers

Few previously studies have examined the HMO degrading abilities of *R. gnavus*, which made it difficult to find specific primers for the analysis. The shotgun sequencing data analysis was delayed, so new primers could not be designed based on the sequenced genome. A few primers were therefore found and decided to be sufficient. When tested, only a few gave signals in qPCR when tested on *R. gnavus*. Five of these were chosen for the experiment, varying in function and all being related to HMO and/or mucin degradation.

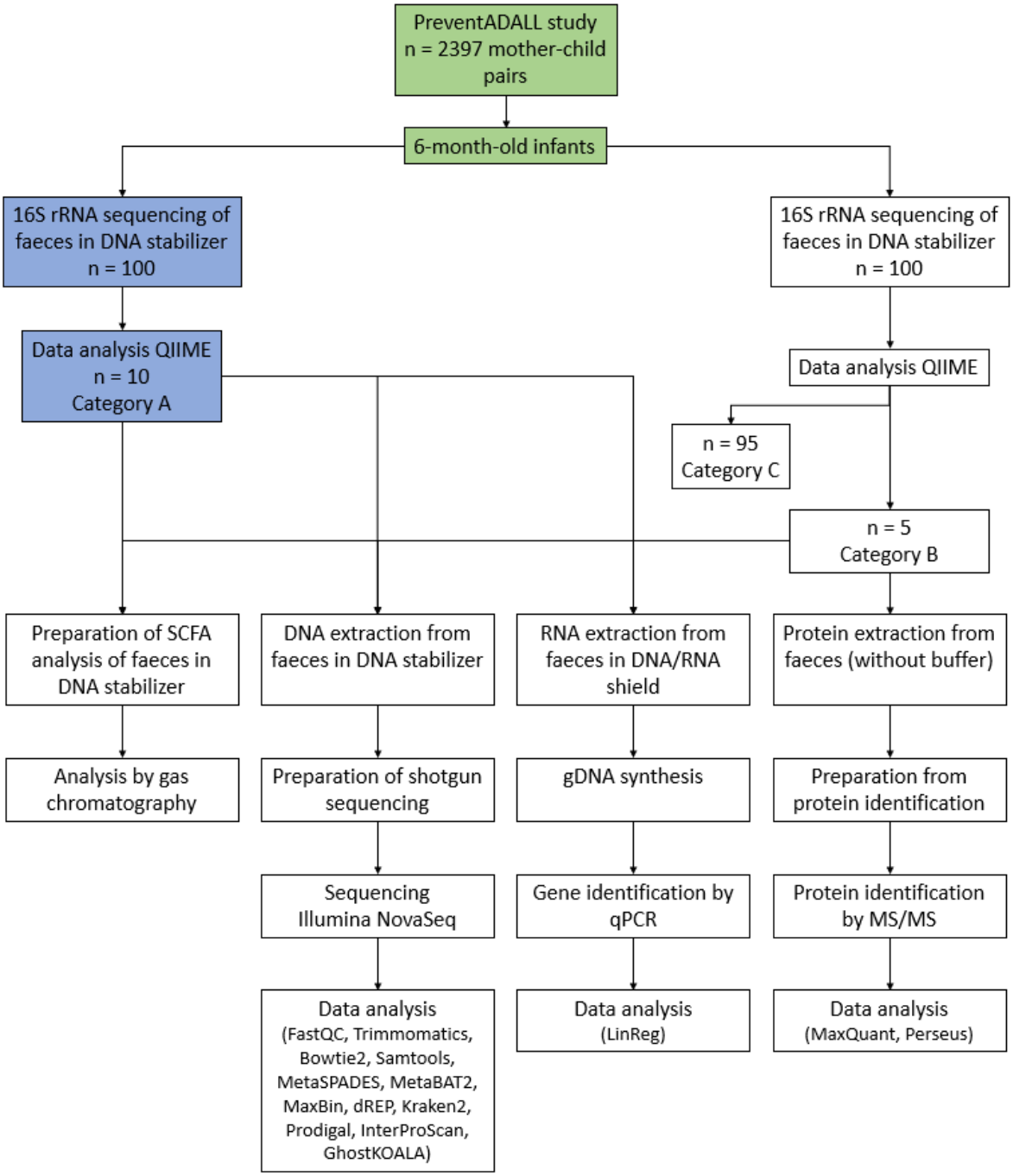

Figure A.1: Workflow overview of master thesis*. Samples were collected through the PreventADALL study, and samples from 6-month-old children were used in this thesis. Green boxes were done previously. 10 samples high in R. gnavus, based in 16S rRNA sequencing by PhD Morten Nilsen, were initially planned to be used for both shotgun sequencing, transcriptomics and proteomics. The 10 samples were not collected without buffer, and proteins could not be extracted. DNA from 100 random samples were therefore extracted, and the 16S rRNA were sequenced. From the 100 samples, 5 samples high in R. gnavus were chosen. Shotgun sequencing were conducted on all 15 samples high in R. gnavus, as well as SCFA analysis by gas chromatography. RNA identification was conducted on 14 samples, as 1 sample was lacking in DNA/RNA shield buffer. Protein analysis using mass spectrometry were conducted on 5 samples high in R. gnavus. 16S rRNA sequencing data was analysed using QIIME pipeline, and shotgun data was analysed using a series of different analysing tools. qPCR results were adjusted using LinReg. Mass spectrometry data was analysed using MaxQuant and Perseus. It was performed PCA analysis and correlation analysis using RStudio.*

Appendix B: Primer sequences

Table B.1*: Primer sequences for PRK primers used for amplification of 16S rRNA gene V3 and V4 region.*

Primer name	Primer sequence (5' -> 3')	Reference
PRK-341F (forward)	CCTACGGGGRBGCASCAG	(Yu et al., 2005)
PRK-806R (revers)	GGACTACYVGGGTATCTAAT	

Table B.2: *The primer sequences of primers used for gene identification by qPCR. The table lists the gene name, forward and revers primer sequences and the activity of the gene product.*

Gene name	Primer sequence - Forward	Primer sequence - Revers	Activity of protein	Reference
RUMGNA_01058	ATCCGGAAAGACCAGACTCC	TTTCCAGACGTCGATCCAAT	Fucose activity	(Crost et al., 2013)
RUMGNA_01638	CCACAGGTTCTTATGTCCGTTT	ATCACCCTTTTTCCGATCAA	β-galactosidase	
RUMGNA_02693	TGCAGGAGTCAAACACAAGG	CCTTGCCTTTTGGGGTGTA	Epimerase (*NanE*)	
RUMGNA_03611	TTCGGAAAACGGGAGGAT	GCTTTTCCAGATTCGGATACC	α-galactosidase	
RUMGNA_03833	CCAATTACGGAAAGCTGGAT	TCTGCTTTCCATGTATCTTCACA	α-L-fucosidase	

Table B.3: *Sequences of PRK index primers used for 16S rRNA gene sequencing. The table lists 16 forward primers (F) and seven revers primers (R).*

Primer name	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

Table B.4: *Table lists the sequences of adapter primers used for shotgun sequencing.*

Primer name	**Primer sequence (5’ -> 3’)**	**Index**
H503	TATCCTCT	Index 1 (i5)
H505	GTAAGGAG	Index 1 (i5)
H506	ACTGCATA	Index 1 (i5)
H517	GCGTAAGA	Index 1 (i5)
H705	AGGAGTCC	Index 2 (i7)
H706	CATGCCTA	Index 2 (i7)
H707	GTAGAGAG	Index 2 (i7)
H710	CAGCCTCG	Index 2 (i7)
H711	TGCCTCTT	Index 2 (i7)
H714	TCATGAGC	Index 2 (i7)

Appendix C: R scripts

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

df <- data.frame(matrix(ncol = 2, dimnames = list(NULL, c("Seq.ID",
"Bin_ID"))))
bin <- list.files(path = "Bins/", pattern = "*.fa", full.names = T)
for (i in 1:length(bin)) {
  lines <- readLines(bin[i])
  logicals <- str_detect(lines, pattern = ">")
  idx <- which(logicals)
  fasta.table <- tibble(Seq.ID = lines[idx])
  fasta.table$Bin_ID <- rep(bin[i], nrow(fasta.table))
  df <- rbind.data.frame(df, fasta.table)
}
data <- df[-1,]
data$Seq.ID <- gsub(pattern = ">", replacement = "", data$Seq.ID)
new_krk.tbl1 <- arrange(new_krk.tbl, Seq.ID)
data1 <- arrange(data, Seq.ID)
tabell <- bind_cols(new_krk.tbl1, data1)
tabell <- tabell[, -6]

library(dplyr)
gnavus <- tabell %>% filter(str_detect(Tax.ID, "\\gnavus")) %>%
filter(str_detect(Bin_ID, "Bins/Sample", negate = T)) %>%
filter(str_detect(Bin_ID, "Bins/T", negate = T))
gnavus1 <- arrange(gnavus, Bin_ID)

df2 <- data.frame(matrix(ncol = 3, dimnames = list(NULL, c("Seq.ID",
"AA.sequence", "Bin_ID"))))
aa <- list.files(path = "AA/", pattern = "*.fa", full.names = T)
for (k in 1:length(aa)) {
  lines2 <- readLines(aa[k])
  idx2 <- which(str_detect(lines2, pattern = ">"))
  fasta.table2 <- tibble(Seq.ID = lines2[idx2])
  N.rows <- nrow(fasta.table2)
  for (row in 1:N.rows) {
    seq.line.first <- idx2[row] + 1
    if(row == N.rows){
      seq.line.last <- length(lines2)
```

```
    } else {
      seq.line.last <- idx2[row + 1] - 1
    }
    seq.lines <- lines2[seq.line.first:seq.line.last]
    fasta.table2$AA.sequence[row] <- str_c(seq.lines, collapse = "")
  }
  fasta.table2$Bin_ID <- rep(aa[k], nrow(fasta.table2))
  df2 <- rbind.data.frame(df2, fasta.table2)
}
aminosyrer <- df2[-1,]
aminosyrer$Bin_ID <- gsub(pattern = ".faa", replacement = "",
aminosyrer$Bin_ID)
aminosyrer$Seq.ID <- gsub(pattern = ">", replacement = "", aminosyrer$Seq.ID)

gnavus_aminosyre <- data.frame(matrix(ncol = 3, dimnames = list(NULL,
c("Seq.ID...2", "AA.sequence", "Bin_ID"))))
for (l in 1:nrow(gnavus)) {
  a <- grep(gnavus$Seq.ID...2[l], aminosyrer$Seq.ID)
  gnavus_aa <- tibble(Seq.ID...2 = aminosyrer$Seq.ID[a], AA.sequence =
aminosyrer$AA.sequence[a])
  gnavus_aa$Bin_ID <- rep(gnavus$Bin_ID[l], nrow(gnavus_aa))
  gnavus_aminosyre <- rbind.data.frame(gnavus_aminosyre, gnavus_aa)
}
gnavus_aminosyre <- gnavus_aminosyre[-1,]
gnavus_aminosyre <- arrange(gnavus_aminosyre, Bin_ID)
gnavus_aminosyre$AA.sequence <- gsub(pattern = "\\*", replacement = "",
gnavus_aminosyre$AA.sequence)
g <- gnavus_aminosyre %>% group_by(Bin_ID) %>% count(Bin_ID)

library(ampir)
df_to_faa(gnavus_aminosyre, file = "Gnavus_kontigs_aa.fasta")
```

Figure C.1: *R script of the pipeline used to make the database containing amino acid sequences from R. gnavus based on shotgun sequencing.*

```
library(readxl)
library(Hmisc)
library(corrplot)
library(RcmdrMisc)
filtered_taxa <- read_excel("16S_prosent_ny.xlsx")
scfa1 <- read_excel("gc_prosent.xlsx")

tbl1 <- rbind(filtered_taxa, scfa1)
tbl2 <- as.data.frame(t(tbl1))
tbl <- tbl2[-1,]
colnames(tbl) <- tbl2[1,]
tbl <- lapply(tbl, as.numeric)
tbl <- as.data.frame(tbl)

corr_alle <- rcorr.adjust(tbl, type = "spearman")
corrplot(corr_alle$R$r, type = "upper", p.mat = corr_alle$R$P, sig.level =
0.05, insig = "blank", tl.col = "black")
```

Figure C.2: *R script of the correlation analysis, using spearman correlation with FDR correction using Holm's method and significance level of 0.05.*

Appendix D: Mass spectrometry specifications

Peptides were loaded to a trap column (Acclaim PepMap100, C18, 5 µm, 100 Å, 300 µm i.d. x 5 mm) and backflushed to a 50 cm x 70 µm analytical column (Acclaim PepMap RSLC C18, 2 mm, 100 Å, 75 µm i.d. x 50 cm, nanoViper). Proteins were separated at a flow rate of 300 nL/min, using a 120 min gradient from 3.2 to 36% solution B (99.9% CAN, 0.1% formic acid). The Q-Exactive mass spectrometer was set up as a Top5 method with a full scan (300-1600 m/z) at R=70.000. Using NCE setting of 28, there followed (up to) 12 MS2 scans at R=17.500. Precursors with single charges (z), and those with z > 5, were excluded for MS/MS, and the dynamic exclusion was set to 20 seconds.

Appendix E: Gas chromatography 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: 4 mm x 6.3 mm x 78.5 mm (Catalog#23311.5, Restek)

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

Column:

Stabilwax DA 30 m, 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: 175 °C

Hydrogen: 30 mL/min

Air: 300 mL/min

Makeup gas: 30 mL/min

Appendix F: Rarefaction curve

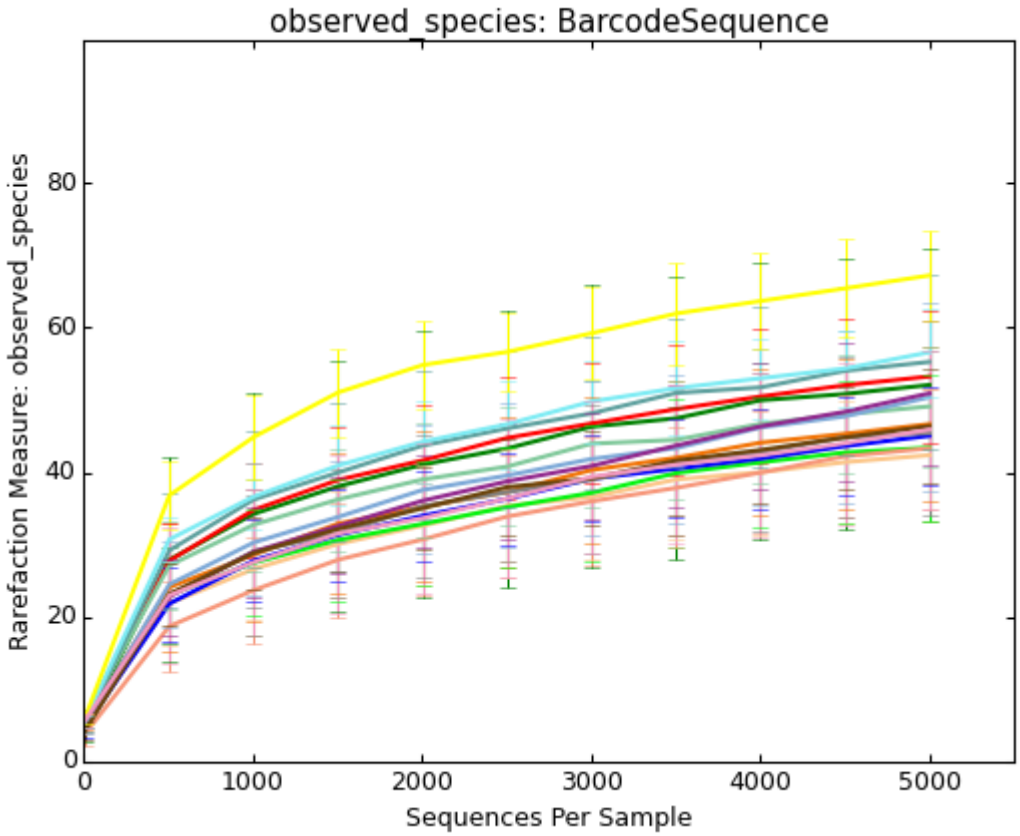

Figure F.1: *Rarefaction curve showing observed species against sequences per sample based on 16S rRNA sequencing results.*

Appendix G: Analysis of SCFA and bacterial composition in the samples

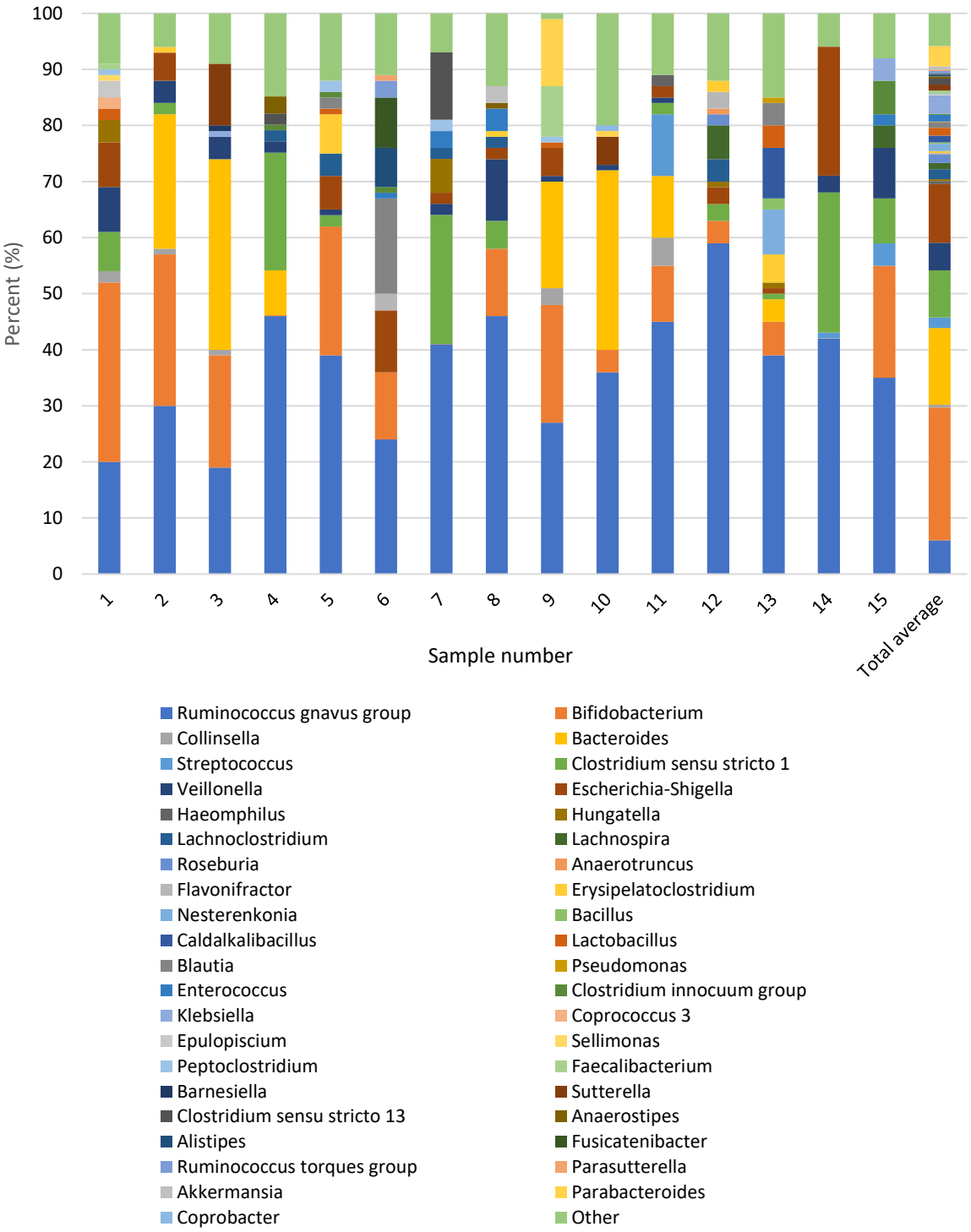

Figure G.1: *Figure shows bacterial composition in all 15 samples, in addition to the average levels in the 95 samples not used for further analysis. Bacterial taxa with higher than 1% abundance in at least one infant are included in the figure. Less abundant taxa are included in "Other".*

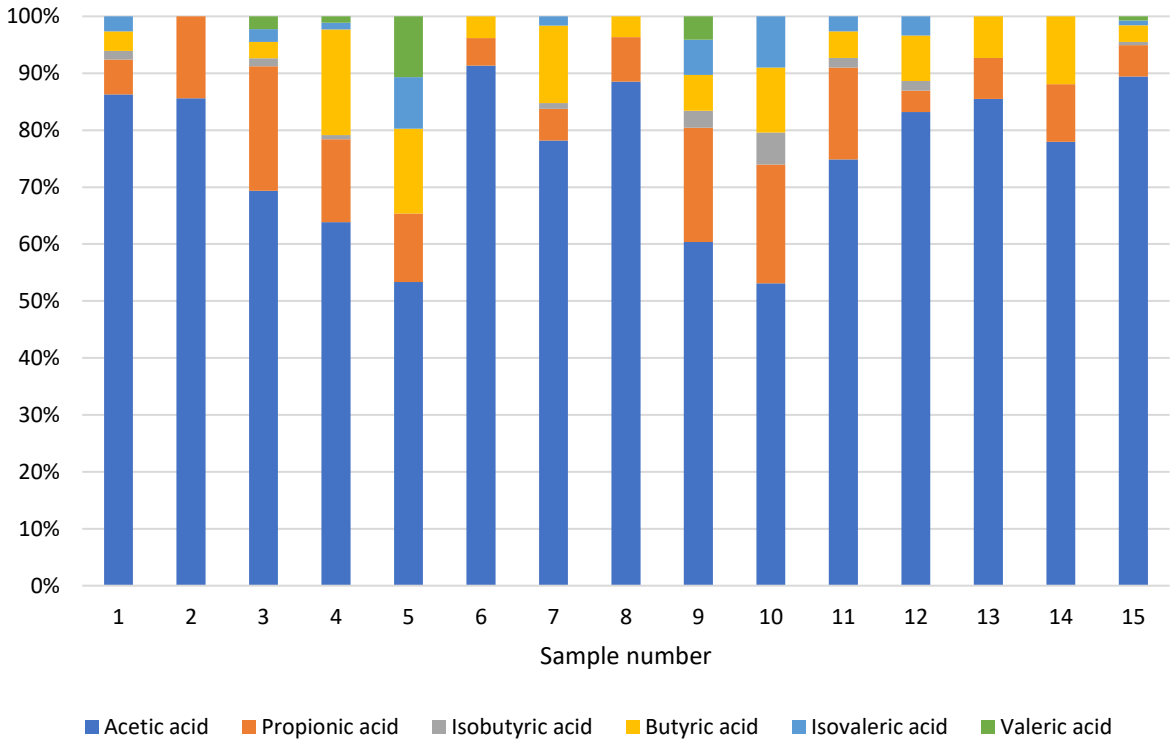

Figure G.2: *Figure shows relative levels of short chain fatty acids in the 15 samples analysed.*

Table G.1: *Table shows p-values for rho in correlation analysis between bacterial taxa and short chain fatty acids. Values marked * are significant (<0.05).*

	Bifidobacterium	*Bacteroides*	*Parabcteroides*	*Enterococcus*	*Streptococcus*	*Clostridium* sensu stricto 1	*Clostridium* sensu stricto 13	*Blautia*	*Hungatella*	*Lachnoclostridium*	*Ruminococcus gnavus*	*Erysipelatoclostridium*	*Veillonella*	*Sutterella*	*Escherichia/Shigella*	Acetic acid	Propionic acid	Isobutyric acid	Butyric acid	Isovaleric acid	Valeric acid
Bifido-bacterium		0.639	0.280	0.929	0.859	0.132	0.366	0.533	0.371	0.621	0.012*	0.533	0.243	0.385	0.969	0.567	0.533	0.814	0.008*	0.595	0.199
Bacteroides			0.253	0.003*	0.032*	0.053	0.867	0.237	0.447	0.101	0.647	0.151	0.989	0.030*	0.308	0.049*	0.0001*	0.151	0.491	0.508	0.728
Para-bcteroides				0.037*	0.345	0.410	0.436	0.616	0.299	0.076	0.017*	0.162	0.353	0.029*	0.418	0.280	0.098	0.043*	0.323	0.021	0.200
Entero-coccus					0.104	0.368	0.635	0.135	0.249	0.044*	0.524	0.773	0.611	0.051	0.771	0.013*	0.010*	0.103	0.819	0.217	0.634
Strepto-coccus						0.139	0.207	0.200	0.187	0.919	0.271	0.581	0.524	0.255	0.507	0.231	0.216	0.216	0.929	0.525	0.928
Clostridium sensu stricto 1							0.178	0.143	0.748	0.567	0.038*	0.393	0.098	0.001*	0.451	0.666	0.211	0.506	0.186	0.516	0.607
Clostridium sensu stricto 13								0.374	0.091	0.887	0.766	0.974	0.604	0.418	0.352	0.660	0.491	0.729	0.918	0.635	0.573
Blautia									0.758	0.321	0.666	0.533	0.081	0.594	0.694	0.593	0.334	0.793	0.800	0.614	0.928
Hungatella										0.235	0.829	0.771	0.460	0.769	0.829	0.229	0.056	0.866	0.714	0.654	0.009*

Valeric acid	Isovaleric acid	Butyric acid	Isobutyric acid	Propionic acid	Acetic acid	*Escherichia/ Shigella*	*Sutterella*	*Veillonella*	*Erysipelato-clostridium*	*Ruminococcus gnavus*	*Lachno-clostridium*
											0.068
										0.196	0.177
									0.232	0.629	0.210
								0.385	0.705	0.027*	0.024*
							0.238	0.657	0.621	0.909	0.584
						0.790	0.550	0.265	0.278	0.685	0.869
					0.001*	0.427	0.133	0.620	0.499	0.491	0.068
				0.190	0.070	0.232	0.022*	0.581	0.736	0.793	0.694
			0.849	0.989	0.024*	0.277	0.109	0.055	0.546	0.027*	0.083
		0.239	0.002*	0.207	0.003*	0.568	0.285	0.197	0.369	0.958	0.474
	0.110	0.650	0.782	0.128	0.064	0.757	0.849	0.916	0.769	0.397	0.591

Appendix H: Protein analysis

Table H.1: *Protein concentration in the samples after purification and preparation. Concentration is measured using Nanodrop. Table lists concentration (mg/mL) and absorbance at 205 nm.*

Parallel	Sample nr.	Concentration (mg/mL)	Absorbance (205 nm)
1	11	0.035	1.07
1	12	0.057	1.75
1	13	0.058	1.81
1	14	0.127	3.94
1	15	0.040	1.25
2	11	0.066	2.05
2	12	0.022	0.68
2	13	0.042	1.30
2	14	0.055	1.71
2	15	0.030	0.94

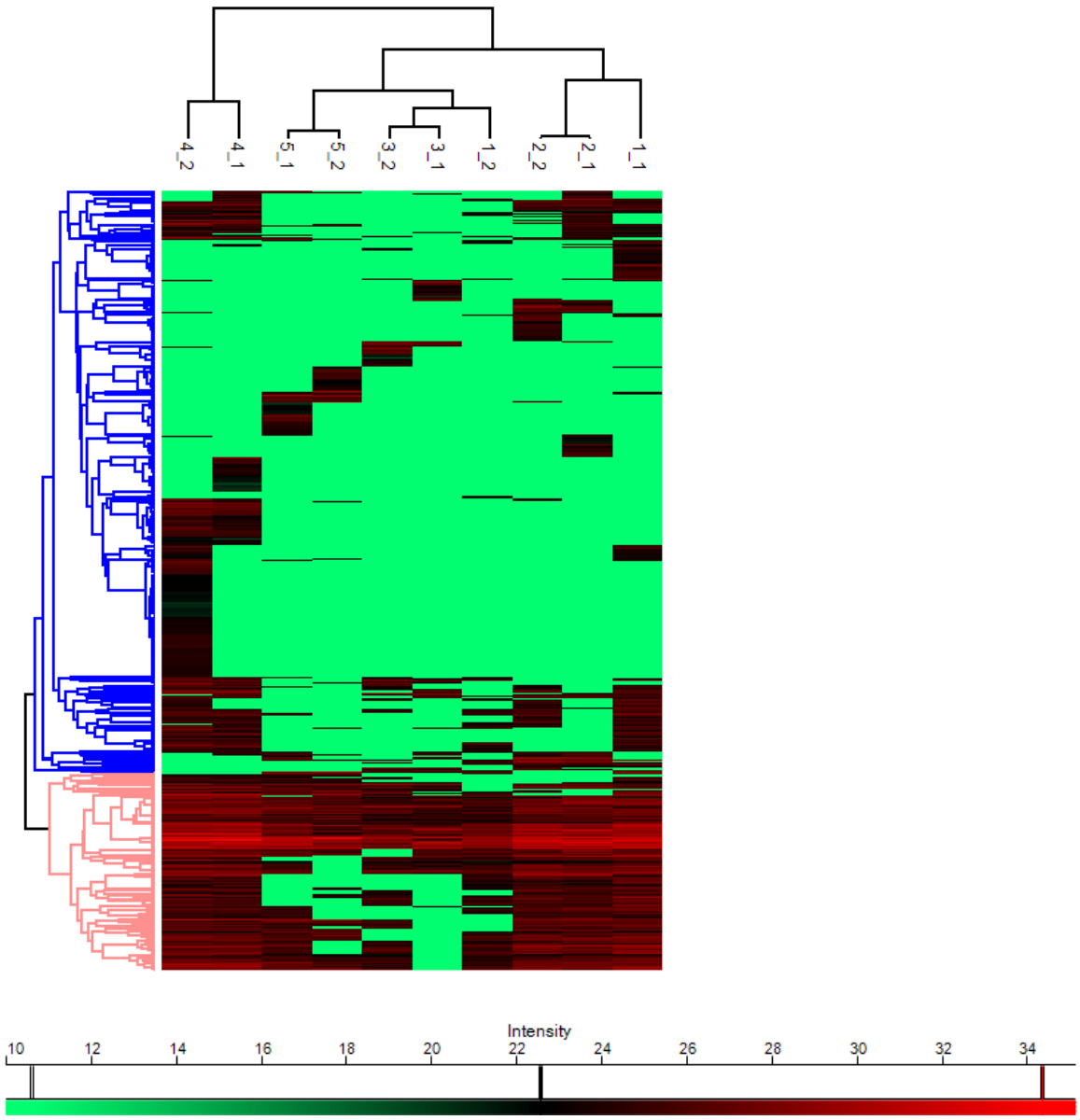

Figure H.1: *Figure shows the result of hierarchical clustering of protein samples using Perseus. Showing two clusters (pink and blue), where the proteins in the pink cluster are most abundant in all samples. The intensity of the bands represents the protein counts, meaning bright red represents the highest abundance.*

Appendix I: Searching for IT-sialidase

One sialidase was identified in the MS data, where proteins were annotated using InterProScan. The sequence ID of the amino acid sequence predicted from shotgun sequencing data matching the amino acid sequence deduced from MS were used to extract the amino acid sequence from the self-constructed database. The amino acid sequence were blasted against the bacteria database on UniProt (BLOSUM-62). A 97% identity was found with the IT-sialidase studied by Crost et al. (2016) and Tailford et al. (2015).

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

Postboks 5003
NO-1432 Ås
Norway