COMPARATIVE GENOMICS AND FUNCTIONAL ANALYSIS OF BIFIDOBACTERIUM CARBOHYDRATE ACTIVE ENZYMES IN HUMAN MILK OLIGOSACCHARIDES UTILIZATION

COMPARATIVE GENOMICS AND FUNCTIONAL ANALYSIS OF BIFIDOBACTERIUM CARBOHYDRATE ACTIVE ENZYMES IN HUMAN MILK OLIGOSACCHARIDES UTILIZATION

By

Grace Kim

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Master of Science

McMaster University

MASTER OF SCIENCE (2024)

Hamilton, Ontario (Department of Biochemistry & Biomedical Sciences)

TITLE: COMPARATIVE GENOMICS AND FUNCTIONAL ANALYSIS OF BIFIDOBACTERIUM CARBOHYDRATE ACTIVE ENZYMES IN HUMAN MILK OLIGOSACCHARIDES UTILIZATION

AUTHOR: Grace Kim

SUPERVISOR: Dr. Michael G Surette

NUMBER OF PAGES: xiii, 111

Lay abstract

The infant gut microbiota is essential for building a beneficial relationship between the gut microbes and the infant. Breastfeeding is associated with a reduced risk of developing asthma. Infants who are not breastfed and are exposed to antibiotics have a higher risk of developing asthma. Previous studies have shown that certain gut bacteria are associated with a reduced risk of asthma. This is because these bacteria have a wide variety of enzymes specialized in utilizing the carbohydrates found in human breast milk. However, we do not fully understand the mechanism behind the protective effects. In this study, I investigated how these bacteria utilize these carbohydrates and found that they have varying capacities to degrade them. I also describe a method to predict and assign the breakdown of carbohydrates in breast milk to the specific genes using computational methods. The breakdown of these carbohydrates varied in these bacteria. This work may help to identify better probiotic strains and carbohydrate supplements for infants that cannot be breastfed.

Abstract

The gut microbiota in early life influences host health and the risk of chronic diseases, including asthma. The infant gut microbiota composition is shaped by multiple determinants such as birth mode, breastmilk feeding practices, and the environment. Previous studies have identified specific taxa associated with protection from atopy and asthma including *Bifidobacterium*. Human Milk Oligosaccharides (HMOs) are an abundant component of breastmilk; however, infants cannot digest them as they lack the necessary enzymes. *Bifidobacterium* produce specific glycoside hydrolases (GHs) that hydrolyze HMOs and release short chain fatty acids, which are beneficial for the infants. The colonization of *B. longum* subsp*. infantis (B. infantis)* in the first year of life is predicted to be protective against asthma development. However, there are other *Bifidobacterium* species or *B. longum* subspecies that colonize the infant gut as well as strain diversity in GHs.

In this study, I used both comparative genomics and phenotypic screening of 118 *Bifidobacterium* strains. Comparative genomics identified strain specific differences in GHs and phenotypic screening showed variability in HMOs degradation among strains. By constructing sequence similarity networks for GHs involved in HMO degradation, I assigned subtypes to GH proteins. These subtypes were hypothesized to have different functions and substrate specificity. Using the machine learning data of the GH subtype profiles combined with HMO utilization assay data, I mapped specific degradation reactions to GH subtypes. Lastly, metagenomic reads were mapped against selected *Bifidobacterium* strains and GH subtype genes. Although *B. infantis* is associated with a reduced asthma risk, I observed that this strain was also abundant in some subjects with an asthma phenotype. Overall, our metagenomic read mapping analysis suggests that asthma development is not solely determined by one *Bifidobacterium* strain or GH subtype. Instead, it appears that multiple factors contribute to asthma risk.

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Michael Surette, for providing me with the privilege to do science with outstanding scientists and become an independent researcher. His support and guidance throughout my graduate school have been invaluable. I still vividly remember the first day I visited the lab, meeting with Mike on a weekend to discuss my project. The moments we spent in his office brainstorming experimental designs on the whiteboard will remain unforgettable. One of the highlights of my master's journey was attending his Murray awardee talk at the CSM, which will forever be invaluable. Mike will always be my mentor and inspiration.

I would like to extend my sincere gratitude to my committee members, Dr. Gerry Wright and Dr. Deborah Sloboda. Your guidance and feedback during committee meetings have been crucial in shaping my academic journey. I am deeply grateful for the lasting impact your mentorship will have on my future.

Thank you to all the Surette lab members for their support. I could not have survived graduate school without every one of you. A special shoutout to Dominique, my favorite scientist, mentor, and friend. She was always there for me, whether going for coffee, getting our daily steps in, helping me with R plots, editing my committee reports, or offering emotional support. She truly made my studies memorable. We shared countless moments of laughter, tears, and celebration. I will miss our Monday gym sessions, and Wednesday carrot muffin and pizza runs. Thank you, Laura, for guiding me through the outline of every experiment I conducted. Though my time with April was short, her positive influence and friendship left a lasting impact. Arsh, thank you for always being early in the lab. I will miss our morning tea sessions and daily walks to the Farncombe office to get water. Sharok, I appreciate all your advice and guidance on bioinformatics. I have become a better researcher and person thanks to you all.

Lastly, I would like to express my appreciation to my friends and family. To Vivian and Alexa, thank you for our remote study sessions, which helped me focus on writing. Yeonjoon, who cried and celebrated many moments with me since elementary school despite the distance and being in different time zones. Thanks to Edith, my younger sister, for her tremendous emotional support, daily snack runs, and encouragement. I am deeply thankful to my parents and grandparents for their unconditional love and support in pursuing graduate school.

Table of Contents

List of Figures

List of Tables

Abbreviations

2’FL	2’-fucosyllactose
3-FL	3-fucosyllactose
3’SL	3’-sialyllactose
6’SL	6’-sialyllactose
BHI	Brain Heart Infusion
CAZyme	Carbohydrate active enzyme
CHILD	Canadian Healthy Infant Longitudinal Development
DFLac	Difucosyllacto-N-tetraose
DSLNH	Disialyllacto-N-hexaose
DFLNT	Difucosyllacto-N-tetraose
DFLNH	Difucosyllacto-N-hexaose
DSLNH	Disialyllacto-N-hexaose
DSLNT	Disialyllacto-N-tetraose
FDSLNH	Fucodisialyllacto-N-hexaose
FLNH	Fucosyllacto-N-hexaose
FOS	Fructo-oligosaccharides

Fuc	Fucose
FUT2	α-1,2-Fucosytransferase
FUT3	α-1,3/4-Fucosyltransferase
Gal	Galactose
GH	Glycoside Hydrolase
Glc	Glucose
GlcNAc	N-acetyl-D-glucosamine
HMOs	Human Milk Oligosaccharides
HPLC	High-Performance fluorescence Liquid Chromatography
iTOL	Interactive Tree of Life
Le	Lewis gene
LNB	Lacto-N-biose
LNFP	Lacto-N-fucopentaose
LNH	Lacto-N-hexaose
LNnT	Lacto-N-neo-tetraose
LNT	Lacto-N-tetraose
LSTb	Sialyl-lacto-N-tetraose b

LSTc	Sialyl-lacto-N-tetraose c
NeuAc	Sialic acid
OD_{600}	Optical density at 600 nm wavelength
PCR	Polymerase chain reaction
pHMOs	Pooled human milk oligosaccharides
PYG	Peptone yeast glucose
RNA-seq	RNA sequencing
SCFA	Short chain fatty acids
Se	Secretor gene
SSN	Sequence Similarity Network

Declaration of Academic Achievement

I, Grace Kim, declare that this thesis titled, “COMPARATIVE GENOMICS AND FUNCTIONAL ANALYSIS OF BIFIDOBACTERIUM CARBOHYDRATE ACTIVE ENZYMES IN HUMAN MILK OLIGOSACCHARIDES UTILIZATION” and the work presented in it are my own, except where noted within each chapter were done in collaboration.

Experimental designs were developed by me and Dr. Michael Surette. Laura Rossi performed library preparation for whole genome sequencing. Dr. Jennifer Stearns provided some of the bifidobacterial strains used in the study. Dr. Lars Bode provided the purified pooled HMOs and his group carried out the glycoprofiling of the HMO utilization assay. Dr. Charisse Peterson provided the CHILD metagenomic data.

In Chapter 2, Dr. Nick Dimonaco assisted with extracting the GH subtype protein sequences and Dominique Tertigas helped process the sequencing data and visualize the carbohydrate utilization data in R.

In Chapter 3, Dr. Nick Dimonaco assisted with implementing WEKA decision tree software and Dr. Shahrokh Shekarriz performed the metagenomic read mapping. Both Dr. Shahrokh Shekarriz and Jake Szamosi provided guidance on the analysis and visualization of metagenomic data.

CHAPTER 1. Introduction

1.1 The human gut microbiome

The human body harbors trillions of microorganisms, and the gut microbiota is the collection of microorganisms in the gastrointestinal tract (Yang et al., 2016). This microbial community, known as the microbiota, comprises bacteria, viruses, fungi, parasites, archaea, and other microorganisms. The term ‘microbiota’ represents the microorganisms, while ‘microbiome’ includes the microbiota, their genetic material, structural components, and metabolites (Berg et al., 2020).

1.1.1 The infant gut microbiome

The infant gut microbiota helps establish a lifelong mutualistic relationship between the gut microbial community and its respective host (Thursby & Juge, 2017). It plays a significant role in early life; it prevents pathogen colonization, produces vitamins and amino acids, and helps maintain the integrity of the gut epithelial cells (Ahearn-Ford et al., 2022). According to the sterile womb paradigm, the maturation of the infant gut microbiota is believed to start at birth (Milani et al., 2017). However, this dogma is challenged by the *in utero* colonization hypothesis, which proposes that gut colonization begins before birth through contact with a placental microbiome (Perez-Muñoz et al., 2017). However, the research supporting *in utero* colonization hypothesis is limited due

to the lack of contamination controls and molecular methods that lack the sensitivity to detect microorganisms with low biomass (Perez-Muñoz et al., 2017). Kennedy et al. (2023) reviewed recent studies that described microbial populations in human fetuses and concluded that microbial signals detected in fetal microbial populations are likely due to contamination during fetal sample collection or the processes of DNA extraction and sequencing (Kennedy et al., 2021).

At birth, gut maturation is influenced by the mode of delivery and gestational age (Milani et al., 2017). Then, it is influenced by many factors, including feeding regime, environment, maternal diet, and lifestyle (Lordan et al., 2024; Sarkar et al., 2021). Vaginally delivered infants developed bacterial communities like the mother's fecal and vaginal microbiota consisting of *Lactobacillus* and *Prevotella*. Stearns et al. (2017) investigated how bacterial communities develop in infants born vaginally with no antibiotic exposure compared to those exposed to antibiotic prophylaxis (IAP) for Group B *Streptococcus* (GBS). They found that IAP for GBS was associated with delayed expansion of *Bifidobacterium* and a persistence of *Escherichia*. Additionally, longer durations of IAP exposure led to a greater delay in the maturation of the microbial community. On the other hand, infants born via Caesarean section (C-section) displayed a less diverse gut microbiota, characterized by lower levels of *Bacteroides*, *Escherichia* and *Bifidobacteria* and an increase in genera

Enterobacteriaceae and *Clostridium* during the first 12 weeks of life (Stearns et al., 2017).

Feeding mode is another major factor influencing early life microbial colonization. Breastfed infants were mostly dominated by *Bifidobacterium* and *Bacteroides*, which efficiently utilize human milk oligosaccharides in the breast milk (Harmsen et al., 2000; Lv et al., 2022). In contrast, formula-fed infants displayed a more diverse microbiota with a higher abundance of *C. difficile* and *Escherichia coli* than breastfed infants (Shaw et al., 2020). Colonization of *C. difficile* is associated with a higher risk of developing atopic outcomes, including eczema, recurrent wheezing, and atopic sensitization (Penders et al., 2007). Following weaning and the introduction of solid foods, the infant gut microbiota begins to resemble an adult-like microbiome primarily dominated by *Firmicutes* and *Bacteroidetes*. Diet remains crucial in driving changes in the microbiome's composition and diversity over the first three years (Koenig et al., 2011). However, these changes in the infant gut microbiota are not limited to microbial compositional change but also involve alterations in the concentration of metabolites such as short-chain fatty acids (SCFAs) (Trompette et al., 2014).

1.2 Human Breast Milk

Breast milk provides ideal nutrition for the infant and is rich in macronutrients (e.g., proteins, carbohydrates, vitamins, and fats), minerals, hormones, cytokines, and growth factors. Many factors influence its composition

and vary by ethnicity, mother's age, lactation stage, and among different individuals (Ballard & Morrow, 2013; Han et al., 2021).

1.2.1 Human Milk Oligosaccharides

Breast milk contains many beneficial substances, including human milk oligosaccharides (HMOs) (Boix-Amorós et al., 2019). HMOs contain a lactose core and are built from five monosaccharides: glucose (Glc), galactose (Gal), N-acetylglucosamine (GlcNAc), fucose (Fuc), and the sialic acid (NeuAc) (**Figure *1.1***). The lactose core can be elongated with repeats of lacto-N-biose type 1 (LNB; Galβ1-3GlcNAc) or N-acetyllactosamine (LacNAc; Galβ1-4GlcNAc), sialylated and/or fucosylated, resulting in over 200 HMO structures identified to date, which are a result of elongation of 19 core structures (Spicer et al., 2022; Zhang et al., 2021). HMOs can reach 20-25 g/L concentrations in human colostrum, which decreases to 5-10 g/L in mature breast milk (Bode, 2012).

The composition of HMOs varies among women, and interpersonal variations are associated with secretor status and Lewis blood type (Bode, 2015). Fucosylation of HMO is determined by two enzymes, α1-2-fucosyltransferase (FUT2) and α1-3/4-fucosyltransferase (FUT3). The FUT2 enzyme, encoded by the secretor gene (*Se*), adds fucose via α1-2 linkages. In contrast, the FUT3 enzyme, encoded by the Lewis blood group (*Le*), adds fucose via α1-3/4 linkages (Bode, 2015; Han et al., 2021; Spicer et al., 2022). HMOs are indigestible by infants themselves due to the lack of enzymes to cleave the glycosidic linkage.

Instead, they are digested by Carbohydrate-Active enZYmes (CAZymes) expressed by microbes in the gut (Thomson et al., 2018).

1.2.2 Functions of Human Milk Oligosaccharides

HMOs have been demonstrated to play many roles, including acting as prebiotics, preventing pathogen infections, and modulating epithelial cell responses (Bode, 2012). Prebiotics are “selectively fermented substances that induce specific changes in the composition and/or activity of the gastrointestinal microbiota, leading to improvements in the host's health and well-being” (Pandey et al., 2015). HMOs are often considered bifidogenic, meaning they selectively stimulate the growth of bifidobacteria, although only specific bifidobacterial strains can efficiently utilize HMOs (Milani et al., 2017). Bifidobacterial strains are predominant colonizers in breastfed infants due to their ability to metabolize HMOs efficiently. Previous genomic analyses have revealed that bifidobacterial strains from infants possess a broad set of genes for carbohydrate utilization, including genes that code for ATP-Binding cassette (ABC) transporters, and CAZymes like and glycoside hydrolases (GHs) and carbohydrate-binding proteins (Milani et al., 2016; Thomson et al., 2018). However, HMO-degrading genes are limited to a few *Bifidobacterium* strains and are not present in all *Bifidobacterium* species. *Bifidobacterium longum* subsp. *infantis* (*B. infantis*) is known to be the most efficient HMO utilizer, possessing a gene cluster for importing and processing HMOs, including four GHs, solute binding proteins, and ABC transporters facilitating the HMO utilization (Sela et al., 2008).

Although many studies focus on bifidobacteria-HMO interactions, some microbes from other taxa also utilize HMOs. Certain *Bacteroides* species including *Bacteroides fragilis, Bacteroides ovatus, Bacteroides thetaiotaomicron, Bacteroides stercorsis, Bacteroides caccae, and Bacteroides vulgatus*, have the ability to metabolize HMO as their sole carbon source (Marcobal et al., 2010, 2011). Additionally, a recent study has shown that *Akkermansia muciniphila* can degrade HMO structures such as 2'-fucosyllactose (2'FL), lacto-N-tetraose (LNT), lactose, and lato-N-triose II (LNT II) *in vitro* in a strain-dependent manner, using key-glycan degrading enzymes and GHs (Kostopoulos et al., 2020; Luna et al., 2022). A unique gene cluster that facilitates the degradation of HMO derivative lacto-N-biose (LNB) was discovered in *Lactobacillus casei* (Bidart et al., 2014). Although these gut microbes have a limited capacity to degrade HMOs compared to *Bifidobacterium* strains, they may still play an essential role in shaping the infant gut microbiota composition.

The protective effects of HMOs can be categorized into two mechanisms. The first involves HMOs exerting selective pressure that enables beneficial bacteria to outcompete other microbes including pathogens, thereby protecting infants from infections (Ackerman et al., 2017). The degradation of HMOs produces organic acids that acidify the environment and inhibit the growth of pathogens (Tan et al., 2014). The second mechanism involves direct interaction with pathogens. Many viral and bacterial pathogens attach to the glycocalyx, a carbohydrate-rich layer attached to the epithelial cells composed of mucins,

glycoproteins, and glycolipids, to infect the host (Argüeso et al., 2021; Kavanaugh et al., 2015). The structural similarity between HMOs and cell surface glycan receptors causes pathogens to bind to HMOs instead of cell surface glycans, blocking their attachment to epithelial cells (Newburg & Grave, 2014). For instance, fucosylated HMOs, such as 2'FL, inhibit the adhesion of *Campylobacter jejuni* (*C. jejuni*), a leading cause of bacterial diarrhea and infant mortality (Ruiz-Palacios et al., 2003). A study of breastfed Mexican infants concluded that lower levels of specific fucosylated HMOs were significantly associated with higher rates of pathogenic diarrhea (Morrow et al., 2004). More specifically, reduced levels of 2'FL were associated with an increased incidence of *C. jejuni*-induced diarrhea. HMOs can also act as antivirals by mimicking histo-blood group antigens (HBGAs), vital for norovirus attachment (Koromyslova et al., 2017).

In addition to mimicking cell surface glycans, HMOs influence pathogen colonization through biofilm formation. Biofilms are clusters of bacteria attached to surfaces and each other, providing increased resistance to antimicrobials (Vestby et al., 2020). A study investigating HMOs as antibiofilm agents found their antibiofilm activity against *Streptococcus agalactiae* (Group B *Streptococcus*) by quantifying biofilm and analyzing structural differences (Ackerman, Doster et al., 2017). While the mechanism behind this activity remains unclear, it has been hypothesized that the external sugars provide the bacterial community with abundant nutrients, reducing the need for biofilm formation. Alternatively, HMOs

might disrupt the bacterial communication pathways essential for biofilm formation (Ackerman, Doster, et al., 2017; Lin et al., 2017)

Lastly, HMOs can directly affect intestinal epithelial and immune cell responses. The glycocalyx serves as a barrier against toxins, and its impairment can lead to gastrointestinal issues. Kong et al. (2019) demonstrated that 2'FL and 3-fucosyllacotse (3-FL) increased the thickness of adsorbed albumin within the glycocalyx of Caco-2 cells. Enhanced albumin adsorption is linked to improved glycocalyx stability and anti-pathogenic effects (Kong et al., 2019). *In vitro* studies have demonstrated that HMO treatments can inhibit cell proliferation and promote increased epithelial differentiation. High doses of 3'-sialyllactose (3′SL) and 6'-sialyllactose (6′SL) significantly decreased cell proliferation in HT-29 and Caco-2 BBe cells (Holscher et al., 2017). HMOs also influence the immune system by either increasing the expression of different Toll-like receptors (TLRs) or inhibiting TLR signaling (Asakuma et al., 2010). For instance, 3'SL and 6'SL are known to elevate levels of both TLR2 and TLR4, whereas lacto-N-fucopentaose I (LNFP I) specifically enhances TLR4 expression. Previous research revealed that 2'FL, 6'SL, and lacto-N-neo-tetraose (LNnT) inhibit TLR5, whereas 3-FL targets TLR5, 7, and 8 in vitro (L. Cheng et al., 2019). These findings highlight how varying HMO combinations can result in diverse immune modulation effects, potentially offering new strategies for disease prevention in infants.

1.2.3 Human Milk Oligosaccharides in infant formula

Breastfeeding is considered the gold standard for infant nutrition, and the World Health Organization (WHO) recommends exclusive breastfeeding for the first six months after birth (Gallier et al., 2015; Walker, 2010). However, following this recommendation can be challenging due to health issues or insufficient milk production by mothers (Oftedal, 2012; Walker, 2010). In this case, infant formula is often used to meet the infant's nutritional needs. Infant formula milk, usually based on cow's milk, aims to replicate the benefits of human breast milk including HMOs, which are unique in breast milk and cannot be found in the milk of other mammals (Hegar et al., 2019).

Infant formulas often contain bioactive agents such as probiotics, prebiotics like HMOs, fructooligosaccharides (FOS) and galactooligosaccharides (GOS), and post-biotics (Fabiano et al., 2021). Previous studies found that infant formula supplemented with HMO shifted the microbiome composition towards that of breastfed infants with higher bifidobacteria (Bosheva et al., 2022). Due to the complexity of synthesizing HMOs, current formulations typically contain simple HMO structures, 2'FL and LNnT (Puccio et al., 2017). Clinical trials have demonstrated that adding HMOs to infant formula is safe, well-tolerated, and supports age-appropriate growth (Marriage et al., 2015; Puccio et al., 2017). When infants were given formula supplemented with 2'FL (0.2 and 1.0g/L) with a caloric density similar to breast milk, no weight, length, or head circumference differences were observed. The absorption profiles of 2'FL in these formulas were

comparable to those in breastfed infants, with 2'FL detected in both plasma and urine (Marriage et al., 2015).

Similarly, research on infant formula supplemented with two HMOs displayed benefits for infant growth, tolerance, and morbidity. This first randomized, controlled clinical trial of formula supplemented with both 2'FL and LNnT reported reduced morbidity, particularly bronchitis, in infants who received supplemented formula (Puccio et al., 2017). However, few studies have investigated the impact of HMO-supplemented infant formulas on infant health. The limited research results in scarce evidence of their potential preventive benefits, highlighting the need for more controlled clinical trials to understand the effects of HMO-supplemented formula better (Fabiano et al., 2021).

1.3 Asthma and breastfeeding

Asthma is a chronic disease that typically develops in childhood, affecting about 334 million people worldwide burdening public health systems (Ahmadizar et al., 2017; Asher & Pearce, 2014; Ferrante & La Grutta, 2018). Risk factors of asthma include wheezing and atopy; however, asthma can be difficult to diagnose as wheezing episodes are challenging to define, and not all wheezing infants develop respiratory disease (Morgan et al., 2005).

Factors present in early life, such as being a preterm infant, having a low birth weight, maternal asthma, and breastfeeding practices, may increase the risk of developing asthma. Many studies have investigated the association between

breastfeeding practices and the risk of asthma development, concluding that breastfeeding is protective against asthma or wheezing disorders (Azad et al., 2017; den Dekker et al., 2016; Oddy et al., 1999; Xue et al., 2021). This protective effect of breastfeeding could be due to its role in preventing infections in the respiratory tract, promoting lung growth, and supporting the maturation of the immune system (Turfkruyer & Verhasselt, 2015; Victora et al., 2016; Waidyatillake et al., 2013). For mothers with asthma, breastfeeding was associated with a reduction in wheezing episodes in their infants (Azad et al., 2017). Exclusive breastfeeding reduced the incidence of wheezing by 62% compared to no breastfeeding, while partial breastfeeding reduced it by 37%. However, breastfeeding supplemented with formula did not display significant protection against wheezing episodes (Azad et al., 2017). Notably, a meta-analysis indicated that the evidence for this association varies between studies, potentially due to differences in study design, study populations, and infant feeding practices that need to be better documented (Xue et al., 2021). Previously, the Canadian Healthy Infant Longitudinal Development (CHILD) Study examined whether infant feeding practices, including breastfeeding and expressed breast milk, are associated with asthma (Arrieta et al., 2015; Dai et al., 2023; Dai et al., 2022). The study found that any infant feeding practices other than direct breastfeeding led to a higher likelihood of possible asthma diagnosis at 3 years (Arrieta et al., 2015). Compared to exclusively breastfed infants, those who received expressed milk had a 43% increased likelihood of asthma

diagnosis, whereas those who were only given formula had a 79% higher likelihood (Klopp et al., 2017).

An increased risk of pediatric asthma development was observed in infants exposed to antibiotics who were not breastfed (Dai et al., 2023; Donovan et al., 2020). A previous metagenomic study has reported that infants exposed to antibiotics without breastfeeding had a 3-fold increased risk of developing asthma than the infants who were breastfed (Dai et al., 2023). Numerous studies have identified specific taxa associated with reduced risk for developing atopy and asthma including *Bifidobacterium* (Akay et al., 2014; Fang et al., 2022). The protective effect of breastfeeding was associated with the enrichment of *B. longum* subspecies *infantis* (Dai et al., 2023).

1.4 Bifidobacteria

Bifidobacteria are Gram-positive anaerobes and dominant in the stools of breastfed infants (Hidalgo-Cantabrana et al., 2017). They belong to the genus *Bifidobacterium*, within the family *Bifidobacteriaceae*, order *Bifidobacteriales* and phylum Actinobacteria. The genus *Bifidobacterium* contains more than 90 species and multiple subspecies (Turroni et al., 2022). Species commonly found in the human gut microbiome contain *Bifidobacterium breve (B. breve), B. longum* subsp*. infantis (B. infantis), B. longum* subsp*. longum (B. longum),* and *Bifidobacterium bifidum (B. bifidum)*, *Bifidobacterium adolescentis (B. adolescentis), Bifidobacterium catenulatum (B. catenulatum),* and *Bifidobacterium pseudocatenulatum (B. pseudocatenulatum)* (Turroni et al., 2012).

1.4.1 Carbohydrate Active Enzymes

Enzymes involved in carbohydrates are known as Carbohydrate-Active enZYmes or CAZymes. Due to the diversity of mono- and polysaccharides, CAZymes are a diverse family of enzymes that degrade complex carbohydrates with high specificity (Cantarel et al., 2009). They have been categorized into different families based on their amino acid sequence similarities, protein folds, and enzymatic mechanisms. As CAZymes classification is based on the similarity of amino acid sequences, it integrates these enzymes' structural and mechanistic features (Henrissat, 1991). The CAZy database covers five different protein domains: Glycoside Hydrolases (GHs), Glycosyl Transferases (GTs), Polysaccharide Lyases (PLs), Carbohydrate Esterases (CEs), and Carbohydrate-Binding Modules (CBMs). Glycoside Hydrolases (GHs) contain glycosidases, which are responsible for the hydrolysis of glycosidic linkages, and transglycosidases, which are responsible for the rearrangement of glycosidic bonds (Cantarel et al., 2009; Henrissat, 1991). Glycosyltransferases (GTs) are responsible for the synthesis of glycosidic linkages (Wiederschain, 2009). Polysaccharide lyases (PLs) perform non-hydrolytic cleavage of glycosidic bonds (Linhardt et al., 1987). Carbohydrate esterases (CEs) hydrolyze carbohydrate esters to facilitate the GHs activity (Armendáriz-Ruiz et al., 2018). Lastly, Carbohydrate-binding modules (CBMs) are non-catalytic proteins that increase the interaction between the enzyme and substrate (Boraston et al., 2004).

1.4.1.1 Glycoside Hydrolases

Glycoside Hydrolases (or glycosyl hydrolases) (EC 3.2.1.x) are essential enzymes to bifidobacteria, enabling them to adapt in the host environment by breaking down complex carbohydrates (Pokusaeva et al., 2011). Unlike the International Union of Biochemistry enzyme nomenclature based on substrate specificity and molecular mechanism, a classification by CAZy is based on amino acid sequence similarities and 3-dimensional structure. Thus, enzymes within the same GH family may have different substrate specificity and modes of action (Van Den Broek et al., 2008). As mentioned above, GHs are a group of enzymes responsible for the hydrolysis of glycosidic linkages in carbohydrates. These enzymes achieve this through two different catalytic mechanisms, retaining and inverting. The retaining mechanism involves a double displacement process, which is an intermediate. In contrast, the inverting mechanism has a single displacement mechanism, resulting in a product with reversed stereochemistry compared to the substrate (Van Den Broek et al., 2008). Moreover, some GHs exhibit transglycosylation activity, catalyzing the transfer of a glycosyl group to form a new glycosidic bond (Qin et al., 2017). One of the GHs exhibiting transglycosylation activity is β-galactosidases, which can produce prebiotics from lactose (Rabiu et al., 2001).

1.4.2 Carbohydrate utilization by bifidobacteria

Simple carbohydrates, including lactose and sucrose, are degraded in the upper gut by the host and other microbes present in the upper gastrointestinal

tract (O'Callaghan & van Sinderen, 2016). Non-digestible carbohydrates, such as complex plant-derived polysaccharides and host-derived carbohydrates (e.g., HMOs), are metabolized in the lower gut where bifidobacteria inhabit (Pokusaeva et al., 2011). The genus *Bifidobacterium* evolved to have one of the highest numbers of genes involved in carbohydrate metabolism among gut commensals (Milani et al., 2016). According to the CAZy classification, the bifidobacterial pangenome is predicted to contain 3385 genes, which include 57 GH families, 13 GT families, and 7 CEs (Milani et al., 2014). Additionally, over 12% of annotated open reading frames of the bifidobacterial genome are predicted to code for enzymes associated with carbohydrate utilization (Milani et al., 2014). The genetic makeup of bifidobacteria for utilizing glycans is often found in glycan-specific gene clusters, which contain enzymes related to transporting sugar, carbohydrate-specific ABC transporters, substrate-binding proteins (SBPs), and GHs (Pokusaeva et al., 2011). Bifidobacteria degrade hexose sugars (e.g., glucose and fructose) through a process called "bifid shunt", where the fructose-6-phosphoketolase enzyme is involved (Pokusaeva et al., 2011). The bifid shunt is an ATP-generating pathway that also produces SCFAs. Theoretically, it yields 2.5 ATP from 1 mole of glucose, along with 1.5 moles of acetate and 1 mole of lactate (Palframan et al., 2003).

Bifidobacterial species' ability to access HMOs is a characteristic limited to bifidobacterial species associated with infants (Alessandri et al., 2021). Many studies have identified gene clusters dedicated to milk carbohydrate degradation.

Based on microarray and functional analysis, *B. breve* is known to possess a gene cluster specialized in utilizing LNT and LNnT (James et al., 2016). In addition, Garrido et al. (2016) discovered the FHMO cluster (Fucosylated Human Oligosaccharide utilization cluster) that contains two fucosidases and genes associated with the import of fucosylated molecules in infant-born *B. longum* strain. However, the genomic arrangement of clusters displays inter- and intra-species variability, and the presence of gene members from these HMO clusters does not consistently lead to bacterial growth in the presence of the given HMOs. Lawson et al. (2020) demonstrated that while *B. breve* strains had essential GHs for fucosylated HMO degradation, but they did not grow on 2'FL. This could be due to the absence of second fucosidase (GH29) or essential transport genes (Lawson et al., 2020).

Bifidobacterial strains associated with breastfed infants use two different HMO degradation strategies (Kim et al., 2013). HMO degradation can occur either intracellularly or extracellularly. During the HMO metabolism by *B. infantis, B. longum,* and *B. breve*, intact HMOs are imported into the cytoplasm through ABC transporters, where intracellular GHs degrade them. Alternatively, for *B. bifidum*, extracellular GHs degrade HMOs into mono- and/or disaccharides, which are then transported into the cell (Thomson et al., 2018). Upon degradation, derivatives of HMO enter the central metabolism pathway (Kim et al., 2013). Metabolism of HMOs leads to the production of SCFAs (specifically lactate, and acetate) (Henrick et al., 2018; Ioannou et al., 2021; Kim et al., 2013). These

SCFAs have many effects, such as acidifying the gut, protecting it from pathogen colonization, and improving intestinal barrier function and immune cell development (Lordan et al., 2024). Acetate produced by *B. infantis* acts as a carbon source for butyrate-producing microbes, illustrating a process known as cross-feeding (Milani et al., 2017). Cross-feeding is often observed within bifidobacterial species or with other bacteria (Xiao et al., 2024). *B. infantis* degrades HMO and produces monosaccharides, lactate and acetate which are further utilized by *Anaerostipes caccae* to produce butyrate (Chia et al., 2021). When *Bifidobacterium* species such as *B. bifidum* and *B. longum* degrade HMOs extracellularly, they release mono- or disaccharides into their surroundings. This promotes cross-feeding among other bifidobacterial species with less efficient HMO utilizing capacity (Xiao et al., 2024).

1.5 Central Paradigm

1.5.1 Central Hypothesis and Objectives

While breastfeeding is associated with the infant gut microbiome and reduced risk of developing asthma, the utilization of HMOs by microbes, metabolites that are produced as a result of HMO degradation, and their relationship with the microbiome remain unknown (Azad et al., 2017; Dai et al., 2023; Klopp et al., 2017; Miliku & Azad, 2018). To address this gap, the CHILD study was designed to investigate the causal role of gut microbiome in childhood asthma (Dai et al., 2022). The CHILD cohort study is a prospective longitudinal

birth cohort study investigating the roles of genetics, genomics, and environment in developing asthma and allergies.

The focus of my work was to identify the specific GH genes in the infant gut microbiome, associated with a decreased risk of developing asthma symptoms. Rather than focusing on species and subspecies, the study aimed to identify specific GH genes responsible for protecting infants against asthma as the capacity to degrade HMOs and produce metabolites is not limited to a single species or subspecies alone. Therefore, I hypothesized that **HMO utilization would be strain-specific, and mapping GH proteins to specific HMO degradation activities coupled with metagenomic data would identify specific protective genes.**

1.5.2 Aims

To address the hypothesis, the research aimed to:

1. Identify strain-specific differences, GH genes, and HMO specificity in *Bifidobacterium* using comparative genomics and functional assays (Chapter 2) and,
2. Using metagenomic data from the CHILD study, investigate the specific gene and pathway associated with protection from asthma (Chapter 3).

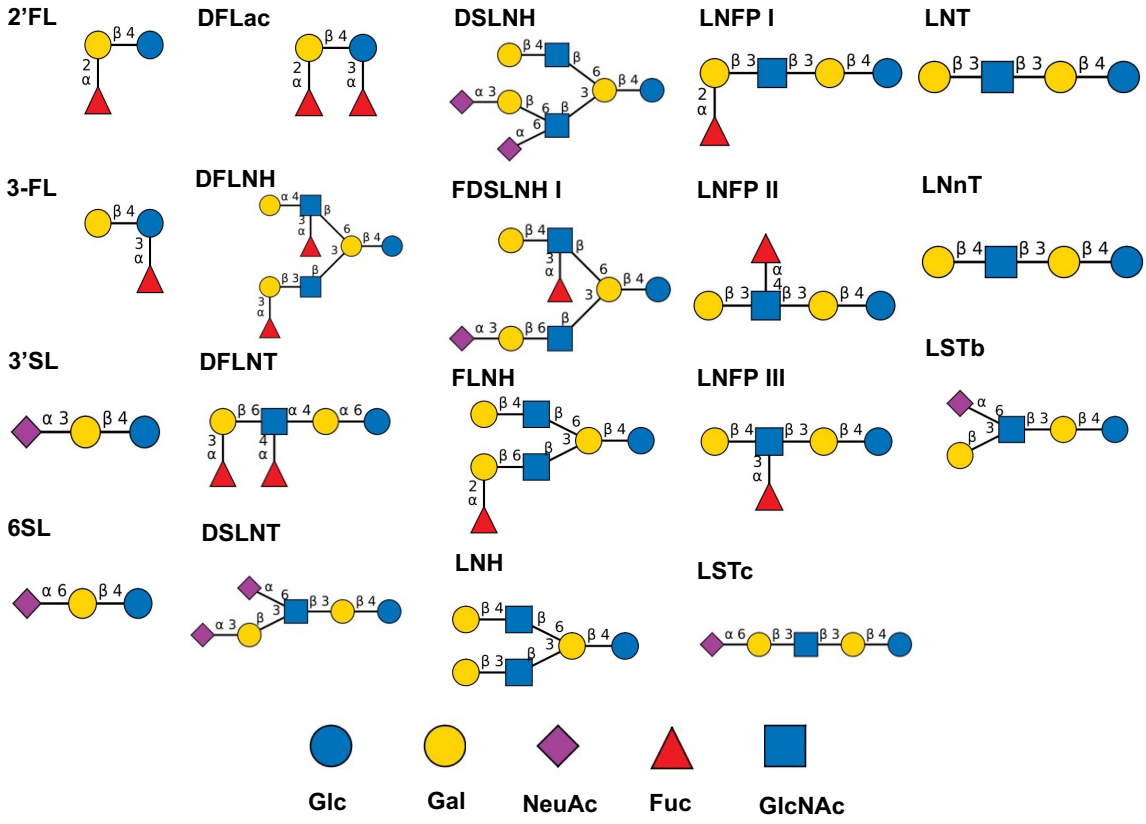


Figure 1.1. Structural diversity of Human Milk Oligosaccharides (HMOs). HMOs are constituted of five different monosaccharides, glucose (blue circle), galactose (yellow circle), N-acetyl-D-glucosamine (blue rectangle), fucose (red triangle), and sialic acid (purple diamond), in various number and linkages, providing high structural diversity. Structures were visualized using DrawGlycan (Cheng et al., 2017).

CHAPTER 2. Strain diversity in Glycoside Hydrolases of *Bifidobacterium* species

2.1 Introduction

Human milk oligosaccharides (HMOs) are an abundant breastmilk component (Boix-Amorós et al., 2019). HMOs in the breastmilk selectively enrich the growth of beneficial microorganisms leading to a healthy gut microbiome (Thomson et al., 2018). However, they cannot be digested by the infants themselves as they lack the enzymes. Previous studies have discovered that many *Bifidobacterium* species have a highly conserved HMO utilization gene cluster allowing preferential utilization of HMO (Garrido et al., 2016; Thomson et al., 2018b). *Bifidobacterium* produces specific glycoside hydrolases (GHs) that can break down HMOs leading to the production of short-chain fatty acids (Janeček & Svensson, 2022; Lordan et al., 2024; Wardman et al., 2022). A previous metagenomic study found a correlation between colonization with *Bifidobacterium longum* subsp. *infantis* (*B. infantis*) in the first year of life and reduced risk of developing asthma, intermediate wheezing, and permanent atopy (Dai et al., 2023). However, infants colonized with *B. infantis* could still develop asthma, while those colonized with other *Bifidobacterium* species may not. Furthermore,

there are numerous *Bifidobacterium sp*ecies and other *Bifidobacterium longum (B. longum)* subspecies that colonize the infant gut, as well as strain-level differences in GHs and capacity to metabolize HMOs (Ioannou et al., 2021a; Zabel et al., 2020). With a long-term goal of identifying the specific set of bifidobacterial genes driving this protective effect against asthma, I have applied comparative genomics and a functional screen to identify GHs and functional differences between *Bifidobacterium* subspecies and strains.

Previous studies have observed substrate specificity of α-fucosidases, a large enzyme family that utilizes fucosylated substrates (Cantarel et al., 2009). For instance, both α-fucosidases (GH29 and GH95) were necessary for *B. breve* to successfully grow on 2-, 3-, and 4-linked fucosylated HMOs (Ruiz-Moyano et al., 2013). However, while the GH95 enzymes showed a preference for 1-2 fucosyl linkages, the GH29 enzymes preferred 1-3 and 1-4 linkages in *B. pseudocatenulatum* strains which highlights the substrate specificity of GH enzymes (Shani et al., 2022). I observed in our isolate collection that one strain can possess multiple genes encoding for members of the same GH family and I predict that the different proteins from the same GH family have different functions and substrate specificities. Therefore, in the present study, I used a bioinformatic approach to investigate whether different GH proteins from the same GH family have different substrate specificities and functions.

2.2 Methods

2.2.1 Sources of bacterial strains and microbial culturing

The Surette lab has a large strain collection of human microbiome isolates including bifidobacteria. All strains were previously isolated and whole genome sequences were available. We restricted our analysis to strains we had available so we could complement comparative genomics with functional studies. A total of 118 strains of *Bifidobacterium* were selected and cultured to create my stock collection (**Table *2.1***). From the frozen stocks in the lab, all strains were grown on Brain Heart Infusion (BHI) agar (Fisher Scientific) supplemented with 1mg/L vitamin K, 10mg/L hemin, and 0.5g/L L-cysteine (BHI3) and incubated anaerobically for 24 hours using an ANANOXOMAT jar (10% H_2, 10% CO_2, 85% N_2) at 37℃. Matrix-Assisted Laser Desorption/Ionization-Time of Flight Mass Spectrometry (MALDI-TOF-MS) on a Bruker Biotyper was used to confirm the identification of the cultured microorganism. Glycerol stocks of each strain were made using pure bacterial colonies on agar, suspended in 1 mL of BHI3 broth with 15% glycerol, aliquoted in each cryovial and two 96 well plates, and stored at -80℃.

To add the diversity of strains, 11 *B. longum* strains isolated from infant stool provided by Dr. Jennifer Stearns were evaluated. They were grown in BHI3 agar, and a single colony was picked for MALDI-ToF identification. The individual spectra from each isolate were compared to remove isolates of the same strain

and four different types of peak patterns were observed. One strain from each peak pattern was selected and whole genome sequencing was carried out using Illumina NextSeq2000. Genomic DNA isolation, library construction, and Illumina sequencing were carried out as described previously with raw read processing (Derakhshani et al., 2020). Assembly, annotation, and taxonomic classification using Unicycler v0.5.0, Bakta v1.5.0 and GTDB-Tk v.2.4.0, respectively (Bolger et al., 2014; Wick et al., 2017; Schwengers et al., 2021; Chaumeil et al., 2020).

2.2.2 Optimizing growth conditions

To identify the optimal growth condition, the strains were cultivated in different base media, methods to minimize the evaporation of the culture, carbon sources, shaking conditions, and different plate types, including 48-well plate and deep 96-well plates.

Using a 96-pin replicator, all bacterial strains in two 96-well plates were transferred to two BHI3 agar and incubated for 48 hours anaerobically in ANANOXOMAT jar as described previously. After 48 hours, 96-pin replicator was used to inoculate four 96-well plates containing 100 μL of BHI3 broth. Two 96-well plates were incubated with a breathable membrane and the other two plates were overlaid with 50 μL of sterile mineral oil. The plates were incubated anaerobically in an ANANOXOMAT jar. The same procedure was used to evaluate the growth in different media, tryptic soy broth (TSB), and peptone-yeast glucose (PYG) broth.

2.2.3 Phenotypic screening under different carbon sources

After optimizing the growth conditions, the 118 strains were cultured in PY broth supplemented with a carbohydrate, autoclaved glucose, or a filter-sterilized carbohydrate including FOS (fructooligosaccharide), N-acetyl-D glucosamine or glucose. Gently, 50 μL of sterile mineral oil was overlaid on the liquid culture in a 96-well plate. The liquid cultures were incubated anaerobically in an ANANOXOMAT jar at 37℃ and after 48 hours, the OD_{600} readings were measured using a microplate reader.

2.2.4 Comparative genomics

Taxonomy and subspecies assignments of 118 *Bifidobacterium* genomes were carried out using GTDB-Tk v.2.4.0 with whole genome assemblies (Chaumeil et al., 2020). Panaroo v.1.4.2 was used to construct a core genome alignment, defined as a set of genes shared by over 95% of strains (Tonkin-Hill et al., 2020). A phylogenetic tree was constructed for each species using FastTree v2.1 and visualized using R v.4.4.0 (Letunic & Bork, 2021; Price et al., 2010). CAZy genes of 8 *Bifidobacterium* species were identified using dBCAN2 (a database of CAZy) (Price et al., 2010; Zhang et al., 2018). The output files of Panaroo (gene presence and absence data) and dbCAN2 were compared to identify the genes associated with carbohydrate metabolism. Based on the presence and absence of the CAZy genes, the unweighted pair group method with arithmetic mean (UPGMA) was used for clustering and visualized in R

v.4.4.0 using tidyverse v.2.0.0, ggtree v.3.12.0 and ape v.5.8 packages for each species (Paradis & Schliep, 2019; Wickham et al., 2019; Yu et al., 2017).

The genus-level phylogenetic trees were constructed by using GTDB-Tk v.2.4.0 for the multiple sequence alignment (MSA) of 120 bacterial marker genes (bac120) identified in the input sequences and FastTree v.2.1 was used to construct a tree of the MSA (Chaumeil et al., 2020; Price et al., 2010). Panaroo was used to construct a genus-level core gene alignment, as described previously, using a protein family sequence identity threshold of 50% and a core threshold of 40% (Tonkin-Hill et al., 2020).

2.2.5 Strains selection

Analysis of the phenotypic screens and comparative genomics was used to remove strain redundancy in this collection. When strains were under the same cluster for both core-gene alignment and CAZy presence and absence tree, their growth under different carbon sources and the sources of the stains were compared. This was done for all 118 strains by species.

2.2.6 Extracting GH protein sequences, building SSNs and identifying GH subtypes

dbCAN2 uses three tools (HMMER, DIAMOND, and eCAMI) to identify CAZy profiles, specifically GH proteins (Zhang et al., 2018). GH protein sequences were extracted for further analysis if predicted by at least two of the three tools as well as all annotations identified using HMMER. For example,

when both HMMER and DIAMOND predicted gene group_3119_1 to be GH13_30, the corresponding protein sequence was extracted. After retrieving all the GH protein sequences, they were categorized by different GH families. The number of GH present in each GH family is shown in **Table *2.3***. Further analysis was limited to the specific GH families known to be involved in HMO degradation (GH2, GH20, GH29, GH33, GH42, GH95, GH112, GH136) (Ioannou et al., 2021b; Saito et al., 2020)

Within each GH family, a protein multiple sequence alignment was performed using MAFFT, and submitted to EFI-EST (Enzyme Function Initiative – Enzyme Similarity Tool, https://efi.igb.illinois.edu/) to generate a SSN (Oberg et al., 2023; Rozewicki et al., 2019; Zallot et al., 2019). The minimum alignment score threshold for drawing edges was determined to be 35%, indicating that nodes that share over 35% sequence identity are connected by edges (Zallot et al., 2019). The percent identity (% id) threshold for nodes varied for each GH protein, established as % id which resulted in the minimum number of nodes and edges. After deciding on the threshold, Cytoscape v.3.8.2 was used to visualize the SSNs to identify clusters within each GH family (Shannon et al., 2003). After visualizing SSNs for GH proteins, protein sequences were submitted to Uniprot to visualize domain structures for each cluster present in each SSN (Bateman et al., 2023). Based on the average protein sequence length and the number of domains, nodes and clusters were assigned GH subtypes and annotated with different colors. Alphabetical labels were used to represent different clusters of

SSNs (e.g. $GH20_A$), and subscripts indicating numbers were used to show variations in protein sequence length and domain structure within the same clusters (e.g. $GH20_{A1}$).

2.3 Results

2.3.1 Comparative genomics and phenotypic screening to capture the strain diversity

Comparative genomics was performed to capture the strain diversity in our collection of 118 *Bifidobacterium* strains (**Table 2.1**). Replicates of the strains isolated from the same source often exist in the lab strain collection. To capture the diversity of the strain collection and to select strains for further analysis, comparative genomics and phenotypic screening were performed.

Phylogenetic trees were constructed based on Panaroo core genome alignment of each species (**Figure 2.1- 2.10**). Carbohydrate-Active enZYmes (CAZymes) are responsible for the synthesis and degradation of polysaccharides. Genes that encode for CAZymes were selected and used to construct a CAZy gene presence and absence tree (**Figure 2.1.A**). A small number of CAZy genes are conserved across all the strains, indicating that a limited set of CAZy genes contributed to the core genome alignment tree. The remaining CAZy genes represent accessory genes that are present in some but not all strains. Additionally, the strains exhibit diversity in CAZy families which contributes to their capabilities of carbon degradation. Two B. adolescentis strains (GC814,

821) are in the same cluster based on the core genome alignment tree, but the CAZy gene content varies (**Figure 2.1.A & B**). Five B. longum strains (GC398, 399, 400, 462, 681, 710) are in the same cluster in the core genome alignment tree (**Figure 2.5.A**) and possess identical CAZy gene presence and absence (**Figure 2.5.B**). In this case, both trees and carbohydrate degradation data were used to select strains.

Isolates were cultured under different carbon sources including autoclaved glucose, filter-sterilized N-acetyl-D-glucosamine, glucose, fructose, or FOS. The capability to degrade carbon sources is strain-specific. The majority of *Bifidobacterium* strains efficiently degraded filter-sterilized glucose; however, not autoclaved glucose. Autoclaving the carbohydrate sources had a negative impact on the growth. N-acetyl-D-glucosamine is degraded poorly compared to other carbon sources, despite being a building block of HMOs. FOS was tested as it is one of the common prebiotics added to infant formula and utilized by most of the *B. longum* strains (Lordan et al., 2024).

Genus-level phylogenetic trees of all *Bifidobacterium* species were constructed using GTDB-Tk (marker gene-based) and Panaroo (core gene-based) (**Error! Reference source not found.**). The phylogenetic tree based on the core genome alignment had a higher resolution than the marker gene-based tree.

2.3.2 Sequence Similarity Network of GHs associated with HMO degradation

The distribution of GH families among *Bifidobacterium* species varies, and it is summarized in Error! Reference source not found.. The heatmap represents the distribution of GH families and the number of genes for each GH family for each genome. The distribution of GH family genes differs between species and within species. Some GH families are common to all isolates (e.g., GH2, 42, 36, 77), whereas some GH families are species-specific (e.g., GH89, 110 in *B. bifidum*). Genes for multiple members of the same GH family also varied between species and strains. For example, genes for the GH43 family were most abundant in *B. scardovii*. It is important to note that not all these GH families are predicted to play a role in HMO degradation.

SSNs were generated, visualized, and categorized for all the GH families that are known to take part in HMO degradation (Error! Reference source not found.Error! Reference source not found.). In a SSN, each node represents a cluster of closely related GH protein and is connected with an edge (Oberg et al., 2023; Zallot et al., 2019). The size of the nodes reflects the number of genes contained within the node. Deciding on the alignment score threshold is important as the network will be fragmented when the alignment score threshold is too high, whereas multiple families will be merged into a single cluster when the alignment score is too low(Oberg et al., 2023; Zallot et al., 2019). Nodes were indicated with colors based on the average protein sequence length within each node, as well

as their domain structures, which were identified using Uniprot. Categorizing SSNs allowed the identification of GH subtypes (e.g. $GH2_{A1}$, $GH2_{A2,}$ $GH2_{B}$). I hypothesize that GH subtypes will have different substrate specificities even if they belong to the same GH family.

For example, the 118 strains contained 361 predicted GH2 proteins. SSN was generated at 55 % identity resulting in a total of 39 nodes and 190 edges (Error! Reference source not found.). I identified 18 subtypes for GH2 proteins. GH2 cluster B ($GH2_{B}$) was categorized not only based on the average protein sequence length but also the number of protein domains. $GH2_{B1}$ had two protein domains whereas $GH2_{B2}$ had one protein domain. Similarly, $GH33_{B2}$ had one protein domain, while $GH33_{B1}$ had two. The node can represent different GH proteins from the same *Bifidobacterium* species or different *Bifidobacterium* species. Our strain collection contains one *B. infantis* strain, however *B. infantis*, has two unique GH2 clusters, $GH2_{A5}$ and $GH2_{G}$.

2.4 Discussion

This chapter had two main objectives. First, comparative genomics and functional assays (growth on different carbohydrates) were used to characterize strains within each *Bifidobacterium* species. Comparative genomics revealed variability in the presence of CAZymes, and GH genes, among strains. These differences in GH gene profiles may have influenced their ability to utilize different carbohydrate sources. For example, FOS, commonly supplemented in infant formulas was only used by a few *Bifidobacterium* species (Fabiano et al., 2021).

Most strains preferred filter-sterilized glucose over other carbohydrate sources and did not utilize N-acetyl-D-glucosamine well despite its role as a building block of HMOs (Kunz et al., 2000). This pattern was observed in most of the *Bifidobacterium* strains in our collection, except a few strains from *B. bifidum*, *B. sp002742445* and *B. dentium.* Strains from these species respectively utilized fructose, FOS, and N-acetyl-D-glucosamine the most. Similar to GH gene profiles, differences in carbohydrate utilization were observed both between species and within species. As described previously, comparative genomics and phenotypic screens were used to select strains. Selected strains were further investigated for their ability to utilize HMOs (Chapter 3).

The second objective was to use sequence similarity networks (SSNs) to compare and subtype GH families across *Bifidobacterium* species and strains. *Bifidobacterium* has an extensive collection of enzymes associated with carbohydrate metabolism among gut commensals (Milani et al., 2016). In our strain collection, 58 GH families were identified, totaling 7296 GHs. According to the previous study on different bifidobacterial genomes, GH13 family members were dominant as *Bifidobacterium* are specialized in breaking down various complex plant polysaccharides (Milani et al., 2016). This was also observed in my study, where the most predominant GH family was GH13, which represents enzymes responsible for hydrolyzing the α-glucosidic linkages in carbohydrates (Bottacini et al., 2018). The second most abundant GH family was GH43, which includes enzymes such as endo-β-xylanases and xylosidases involved in xylan

degradation (Cantarel et al., 2009; Wardman et al., 2022). Interestingly, the *B. infantis* strain AM1522 lacked a representative of GH43. Other highly represented GH families are GH3 and GH42, which include β-galactosidase, an enzyme responsible for degrading lactose (Cantarel et al., 2009; Wardman et al., 2022).

Out of the 58 identified GH families, only 8 known to be associated with HMO utilization were visualized for SSNs. When visualizing SSNs, deciding thresholds for nodes and edges is crucial. A minimum alignment score of 35% for edges is recommended, as annotations between sequences with less than 35% identity are considered unreliable (Zallot et al., 2019). The threshold for a node was determined based on achieving the minimum number of nodes and edges, and it varied for each GH family. At higher percent identities, fewer nodes are observed, and nodes start to separate as the percent identity is decreased. However, this was not always the case. Some nodes remained separate at a high percent identity and remained distinct even when the percent identity was reduced. This suggests that these enzymes may have unique functions within their GH family.

One *Bifidobacterium* strain can possess more than one GH family member, and each GH family catalyzes the hydrolysis of a specific glycosidic linkage. SSNs nodes were categorized into subtypes to determine which glycosidic linkages in each HMO are cleaved by the specific versions of a GH family member. Subtypes were assigned based on both the protein length and the number of protein domains, as protein length could indicate the presence of

another protein domain or carbohydrate-binding modules (Henrissat & Davies, 2000). Domains are structural or functional units within an enzyme, each with a specific role, and their presence often correlates with the overall length of the protein (Marsden & Orengo, 2008). Notable, GH subtypes such as $GH42_{A5}$, $GH2_{G}$, $GH20_{A2}$, $GH33_{A1}$, and $GH33_{B2,}$ were exclusively present in the single *B. infantis* strain in our collection. These GH subtypes were further analyzed in Chapter 3 with HMO utilization assay data and machine learning to assign specific enzymatic reactions to these GH subtypes. Additionally, the protein sequences of these GH subtypes were used to build a reference database for metagenomics read mapping in Chapter 4.

Further investigating these GH subtypes using AlphaFold, a tool that predicts the three-dimensional protein structures, will be valuable. It will provide insights into the substrate specificity of these GHs, structural features, and interactions with other enzymes or substrates, particularly HMOs (Jumper et al., 2021).

A limitation of this study and any comparative genomics study is that the genome assemblies may be incomplete, and some genes may also be missed in the prediction and annotation. Another limitation of this study is that the SSN visualization was restricted to just 8 GH families known to be involved in HMO degradation. However, many enzymes within GH families remain uncharacterized, and their enzymatic functions are unknown (Lombard et al., 2014). For instance, GH20 comprises 23102 enzymes, but only 133 have been

characterized (Lombard et al., 2014). Therefore, expanding the SSN analysis to include all the GH families in our bifidobacterial strain collection would provide a more comprehensive and unbiased view.

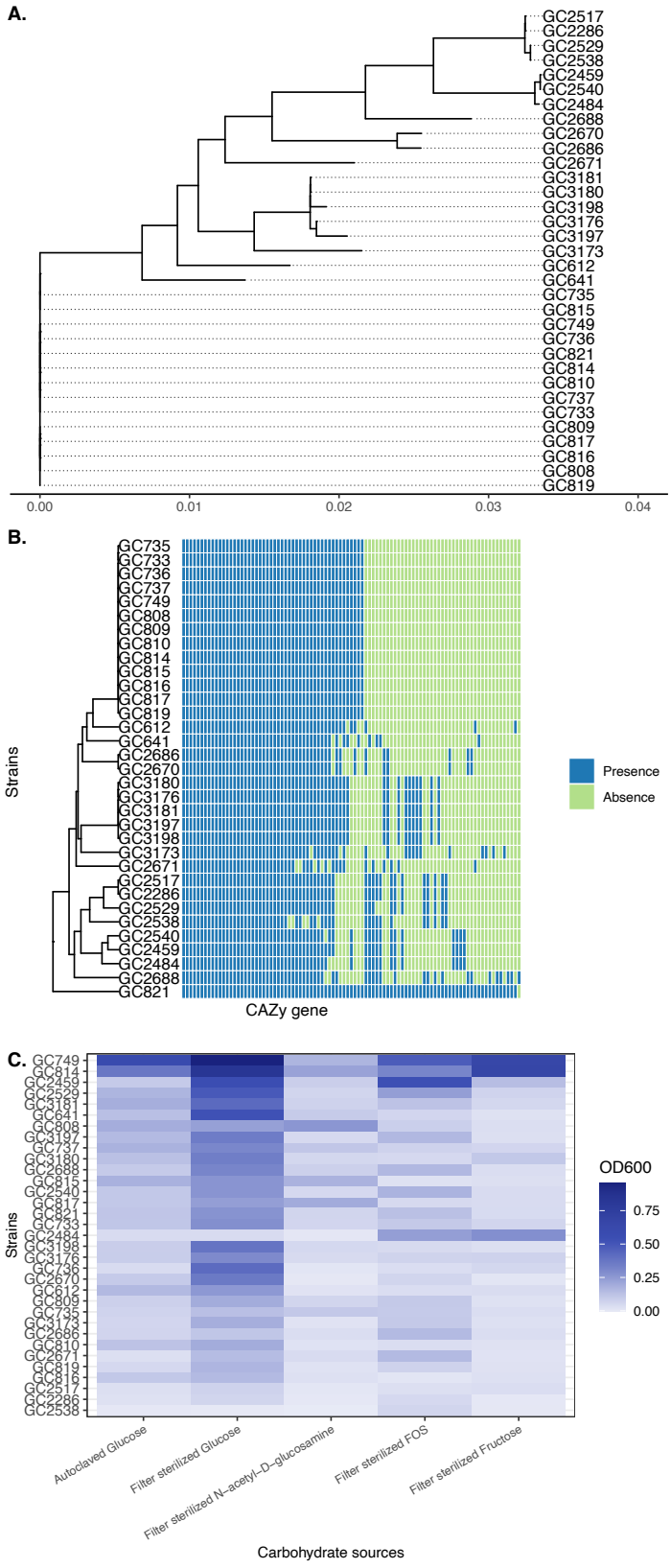

A.
GC2517
GC2286
GC2529
GC2538
GC2459
GC2540
GC2484
GC2688
GC2670
GC2686
GC2671
GC3181
GC3180
GC3198
GC3176
GC3197
GC3173
GC612
GC641
GC735
GC815
GC749
GC736
GC821
GC814
GC810
GC737
GC733
GC809
GC817
GC816
GC808
GC819
0.00
0.01
0.02
0.03
0.04
B.
Strains
CAZy gene
Presence
Absence
C.
Strains
OD600
0.75
0.50
0.25
0.00
Autoclaved Glucose
Filter sterilized Glucose
Filter sterilized N–acetyl–D–glucosamine
Filter sterilized FOS
Filter sterilized Fructose
Carbohydrate sources

Figure 2.1. Comparative genomics and phenotypic screening of *Bifidobacterium adolescentis*. A. A phylogenetic tree based on the core gene alignment generated from Panaroo v.1.4.2 with FastTree v.2.1 (Tonkin-Hill et al., 2020; Price et al., 2010). Taxonomy and subspecies assignments were carried out using GTDB-Tk v.2.4.0 (Chaumeil *et al., 2020).* **B.** Presence and absence of CAZy genes were identified by using Panaroo and dbCAN2 (Tonkin-Hill et al., 2020; Zhang et al., 2018). CAZy gene presence is displayed in blue and absence in green. **C.** Carbohydrate metabolism profile of *B. adolescentis* in PYG media (lacking glucose) supplemented with autoclaved glucose, filter-sterilized fructose, FOS, N-acetyl-D-glucosamine or glucose.

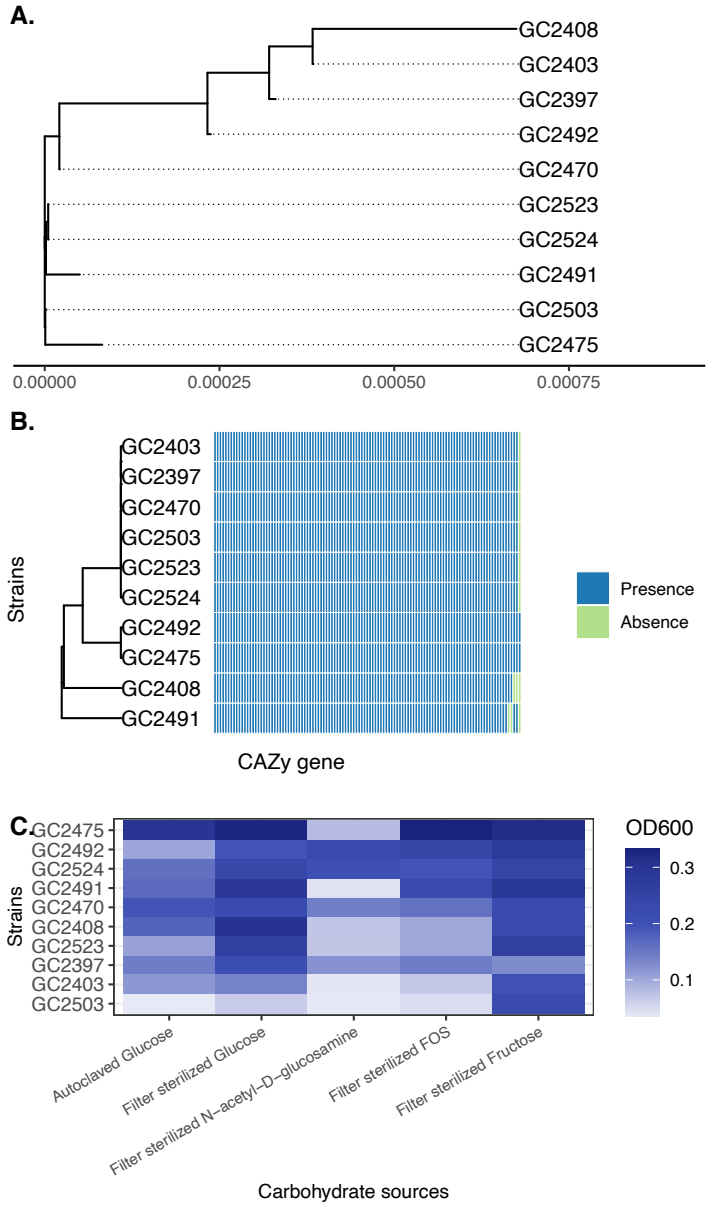


Figure 2.2. Comparative genomics and phenotypic screening of *Bifidobacterium bifidum*. **A.** A phylogenetic tree based on the core gene alignment generated from Panaroo v.1.4.2 with FastTree v.2.1 (Tonkin-Hill et al., 2020; Price et al., 2010). Taxonomy and subspecies assignments were carried out using GTDB-Tk v.2.4.0 (Chaumeil et al., 2020). **B.** Presence and absence of CAZy genes were identified by using Panaroo and dbCAN2 (Tonkin-Hill et al., 2020; Zhang et al., 2018). CAZy gene presence is displayed in blue and absence in green. **C.** Carbohydrate metabolism profile of *B. bifidum* in PYG media (lacking glucose) supplemented with autoclaved glucose, filter-sterilized fructose, FOS, N-acetyl-D-glucosamine or glucose.

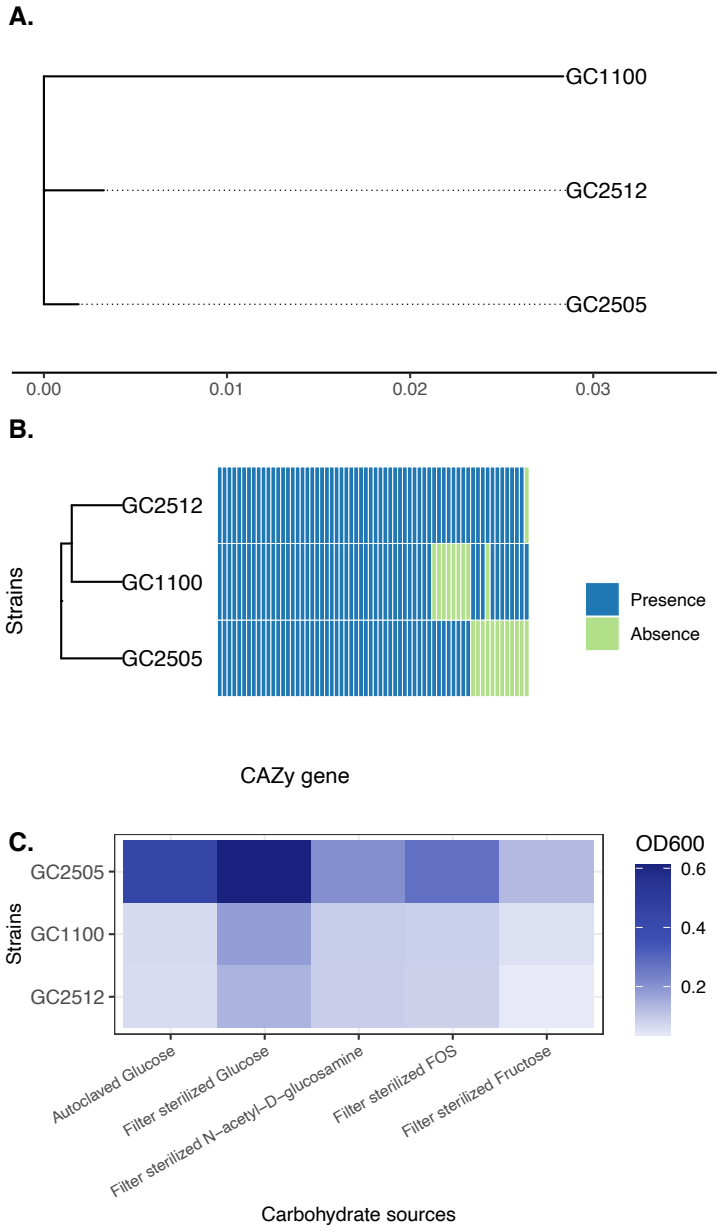


Figure 2.3. Comparative genomics and phenotypic screening of *Bifidobacterium catenulatum*. A. A phylogenetic tree based on the core gene alignment generated from Panaroo v.1.4.2 with FastTree v.2.1 (Tonkin-Hill et al., 2020; Price et al., 2010). Taxonomy and subspecies assignments were carried out using GTDB-Tk v.2.4.0 (Chaumeil et al., 2020). **B.** Presence and absence of CAZy genes were identified by using Panaroo and dbCAN2 (Tonkin-Hill et al., 2020; Zhang et al., 2018). CAZy gene presence is displayed in blue and absence in green. **C.** Carbohydrate metabolism profile of *B. catenulatum* in PYG media (lacking glucose) supplemented with autoclaved glucose, filter-sterilized fructose, FOS, N-acetyl-D-glucosamine or glucose.

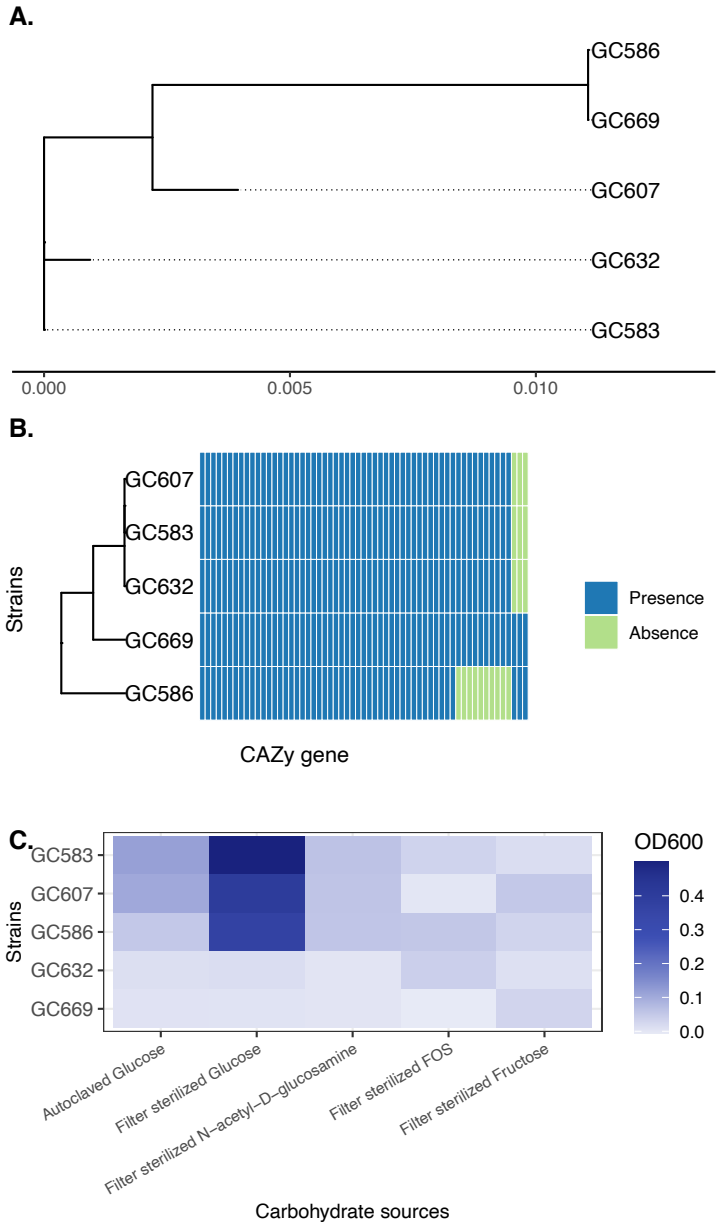


Figure 2.4. Comparative genomics and phenotypic screening of *Bifidobacterium faecale*. A. A phylogenetic tree based on the core gene alignment generated from Panaroo v.1.4.2 with FastTree v.2.1 (Tonkin-Hill et al., 2020; Price et al., 2010). Taxonomy and subspecies assignments were carried out using GTDB-Tk v.2.4.0 (Chaumeil et al., 2020). **B.** Presence and absence of CAZy genes were identified by using Panaroo and dbCAN2 (Tonkin-Hill et al., 2020; Zhang et al., 2018). CAZy gene presence is displayed in blue and absence in green. **C.** Carbohydrate metabolism profile of *B. faecale* in PYG media (lacking glucose) supplemented with autoclaved glucose, filter-sterilized fructose, FOS, N-acetyl-D-glucosamine or glucose.

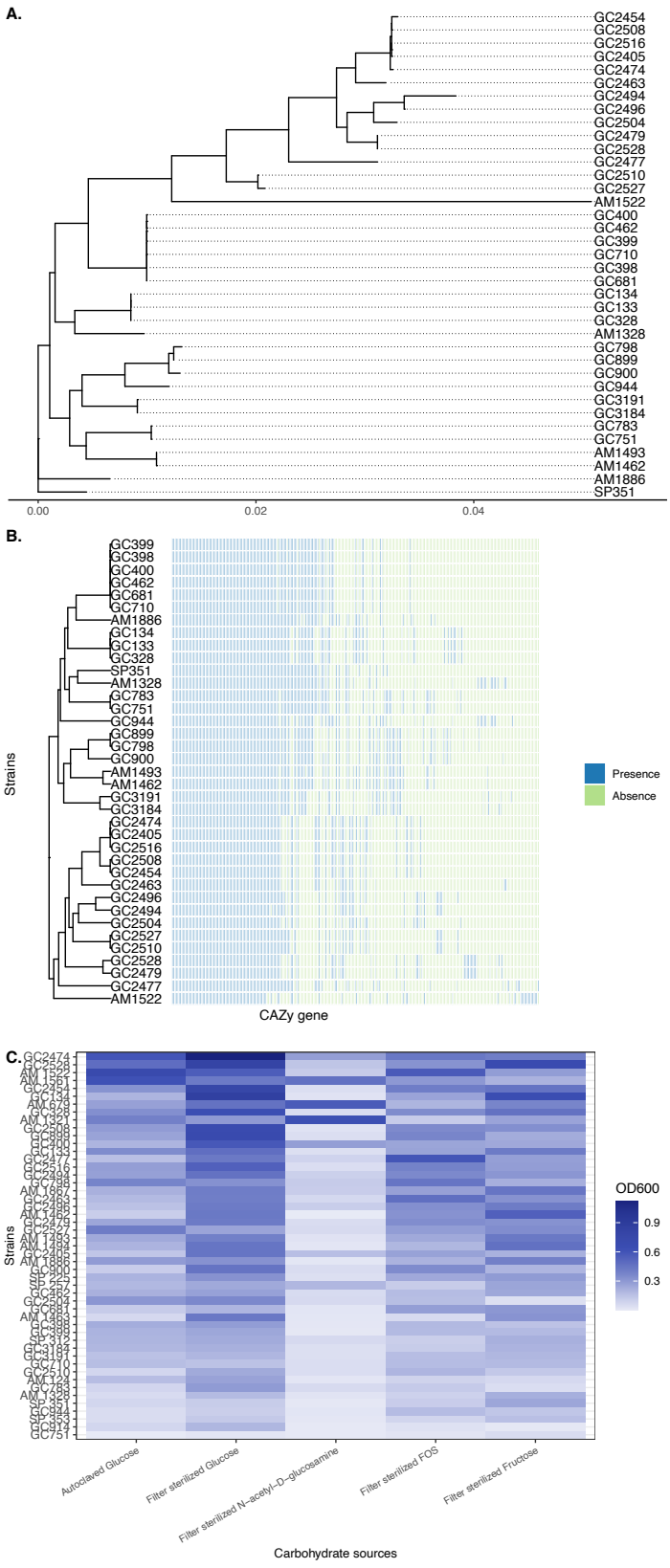

A.
GC2454
GC2508
GC2516
GC2405
GC2474
GC2463
GC2494
GC2496
GC2504
GC2479
GC2528
GC2477
GC2510
GC2527
AM1522
GC400
GC462
GC399
GC710
GC398
GC681
GC134
GC133
GC328
AM1328
GC798
GC899
GC900
GC944
GC3191
GC3184
GC783
GC751
AM1493
AM1462
AM1886
SP351
0.00
0.02
0.04
B.
Strains
CAZy gene
Presence
Absence
C.
Strains
OD600
0.9
0.6
0.3
Autoclaved Glucose
Filter sterilized Glucose
Filter sterilized N-acetyl-D-glucosamine
Filter sterilized FOS
Filter sterilized Fructose
Carbohydrate sources

Figure 2.5. Comparative genomics and phenotypic screening of *Bifidobacterium longum*. A. A phylogenetic tree based on the core gene alignment generated from Panaroo v.1.4.2 with FastTree v.2.1 (Tonkin-Hill et al., 2020; Price et al., 2010). Taxonomy and subspecies assignments were carried out using GTDB-Tk v.2.4.0 (Chaumeil et al., 2020). **B.** Presence and absence of CAZy genes were identified by using Panaroo and dbCAN2 (Tonkin-Hill et al., 2020; Zhang et al., 2018). CAZy gene presence is displayed in blue and absence in green. **C.** Carbohydrate metabolism profile of *B. longum* in PYG media (lacking glucose) supplemented with autoclaved glucose, filter-sterilized fructose, FOS, N-acetyl-D-glucosamine or glucose.

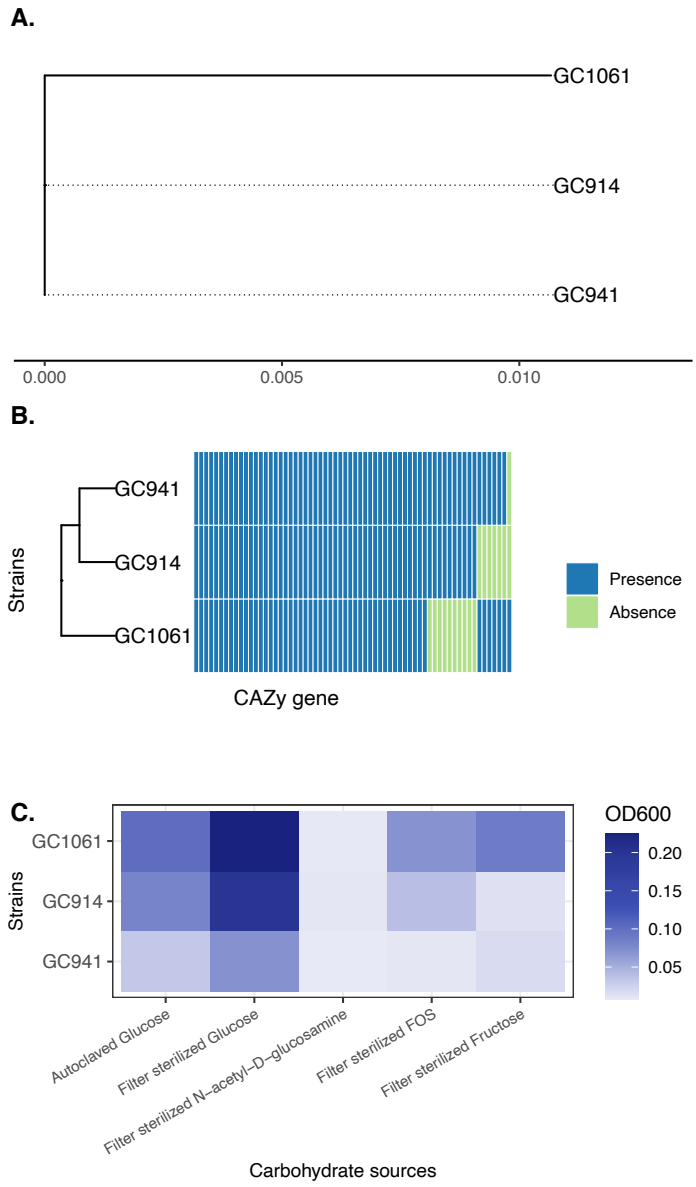


Figure 2.6. Comparative genomics and phenotypic screening of *Bifidobacterium pseudocatenulatum*. A. A phylogenetic tree based on the core gene alignment generated from Panaroo v.1.4.2 with FastTree v.2.1 (Tonkin-Hill et al., 2020; Price et al., 2010). Taxonomy and subspecies assignments were carried out using GTDB-Tk v.2.4.0 (Chaumeil et al., 2020). **B.** Presence and absence of CAZy genes were identified by using Panaroo and dbCAN2 (Tonkin-Hill et al., 2020; Zhang et al., 2018). CAZy gene presence is displayed in blue and absence in green. **C.** Carbohydrate metabolism profile of *B. pseudocatenulatum* in PYG media (lacking glucose) supplemented with autoclaved glucose, filter-sterilized fructose, FOS, N-acetyl-D-glucosamine or glucose.

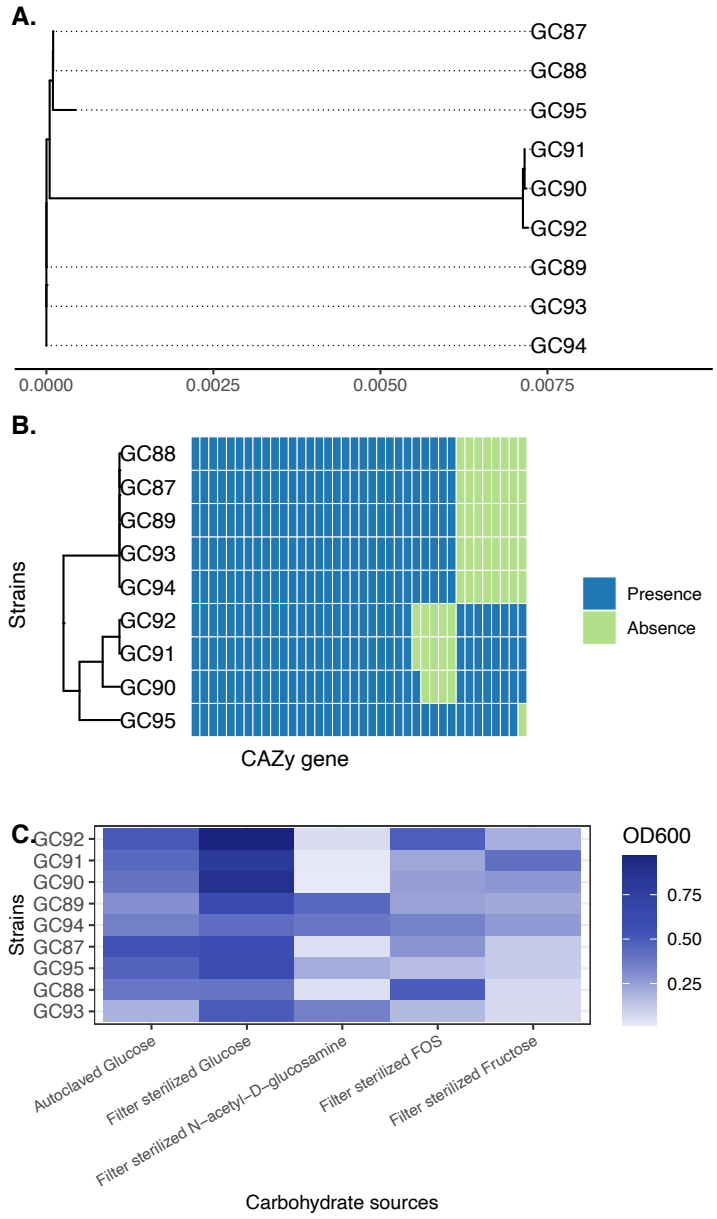


Figure 2.7. Comparative genomics and phenotypic screening of *Bifidobacterium scardovii*. A. A phylogenetic tree based on the core gene alignment generated from Panaroo v.1.4.2 with FastTree v.2.1 (Tonkin-Hill et al., 2020; Price et al., 2010). Taxonomy and subspecies assignments were carried out using GTDB-Tk v.2.4.0 (Chaumeil et al., 2020). **B.** Presence and absence of CAZy genes were identified by using Panaroo and dbCAN2 (Tonkin-Hill et al., 2020; Zhang et al., 2018). CAZy gene presence is displayed in blue and absence in green. **C.** Carbohydrate metabolism profile of *B. scardovii* in PYG media (lacking glucose) supplemented with autoclaved glucose, filter-sterilized fructose, FOS, N-acetyl-D-glucosamine or glucose.

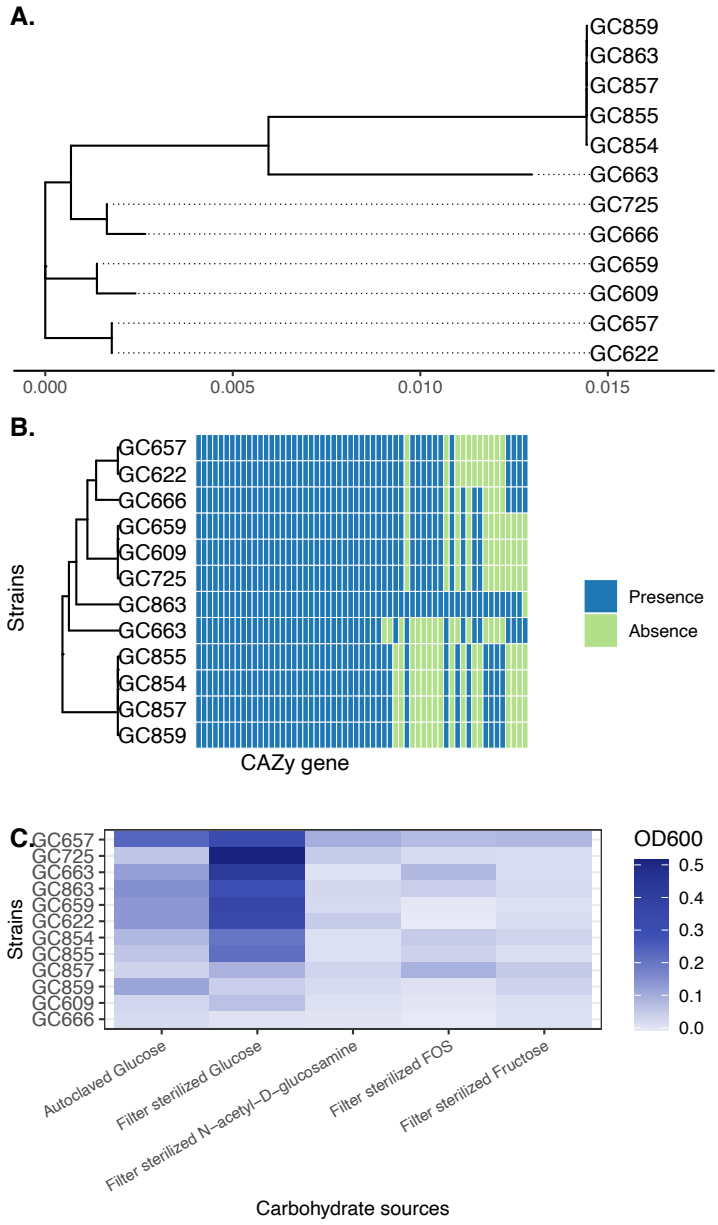


Figure 2.8. Comparative genomics and phenotypic screening of *Bifidobacterium stercorsis.* A. A phylogenetic tree based on the core gene alignment generated from Panaroo v.1.4.2 with FastTree v.2.1 (Tonkin-Hill et al., 2020; Price et al., 2010). Taxonomy and subspecies assignments were carried out using GTDB-Tk v.2.4.0 (Chaumeil et al., 2020). **B.** Presence and absence of CAZy genes were identified by using Panaroo and dbCAN2 (Tonkin-Hill et al., 2020; Zhang et al., 2018). CAZy gene presence is displayed in blue and absence in green. **C.** Carbohydrate metabolism profile of *B. stercorsis* in PYG media (lacking glucose) supplemented with autoclaved glucose, filter-sterilized fructose, FOS, N-acetyl-D-glucosamine or glucose.

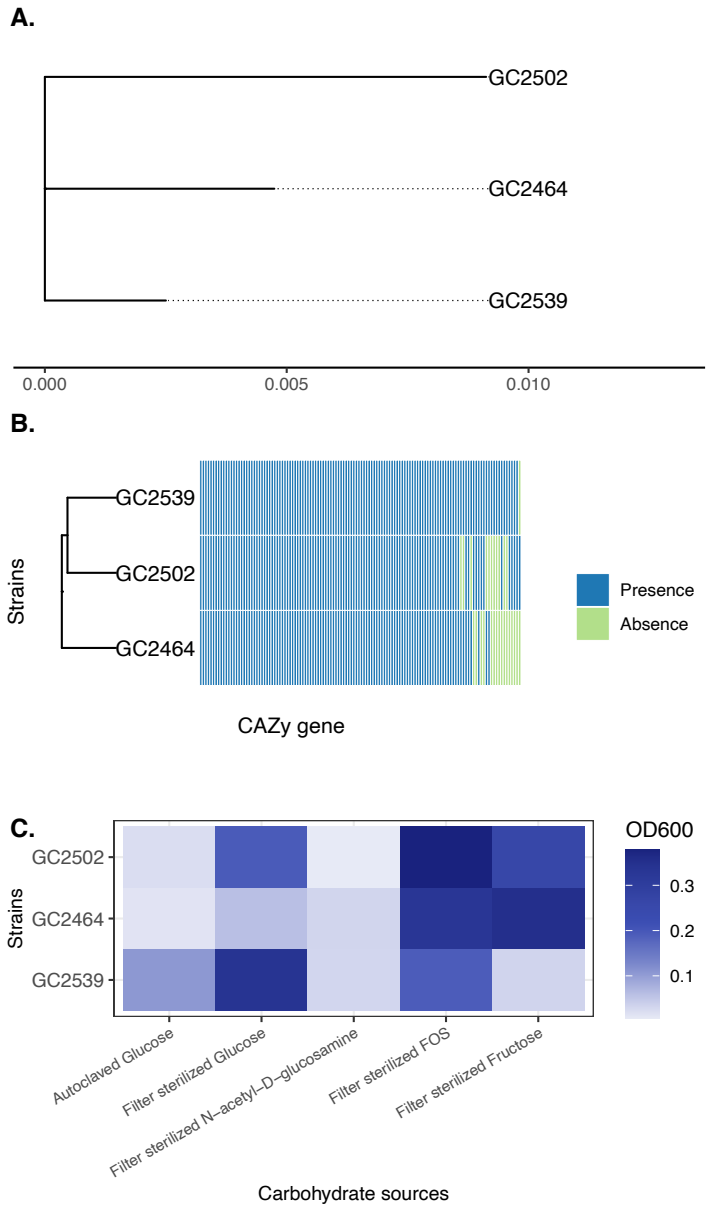


Figure 2.9. Comparative genomics and phenotypic screening of *Bifidobacterium sp002742445*. A. A phylogenetic tree based on the core gene alignment generated from Panaroo v.1.4.2 with FastTree v.2.1 (Tonkin-Hill et al., 2020; Price et al., 2010). Taxonomy and subspecies assignments were carried out using GTDB-Tk v.2.4.0 (Chaumeil et al., 2020). **B.** Presence and absence of CAZy genes were identified by using Panaroo and dbCAN2 (Tonkin-Hill et al., 2020; Zhang et al., 2018). CAZy gene presence is displayed in blue and absence in green. **C.** Carbohydrate metabolism profile of *B. sp002742445* in PYG media (lacking glucose) supplemented with autoclaved glucose, filter-sterilized fructose, FOS, N-acetyl-D-glucosamine or glucose.

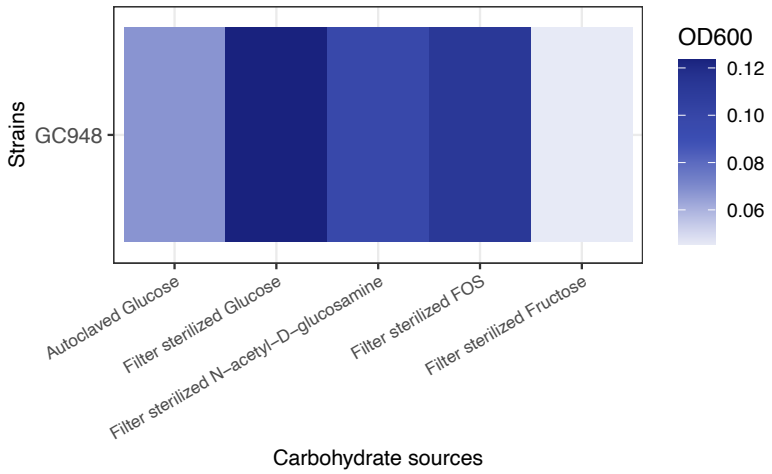


Figure 2.10. Carbohydrate dependent growth of *Bifidobacterium dentium*. Carbohydrate metabolism profile of *B. dentium* in PYG media (lacking glucose) supplemented with autoclaved glucose, filter-sterilized fructose, FOS, N-acetyl-D-glucosamine or glucose.

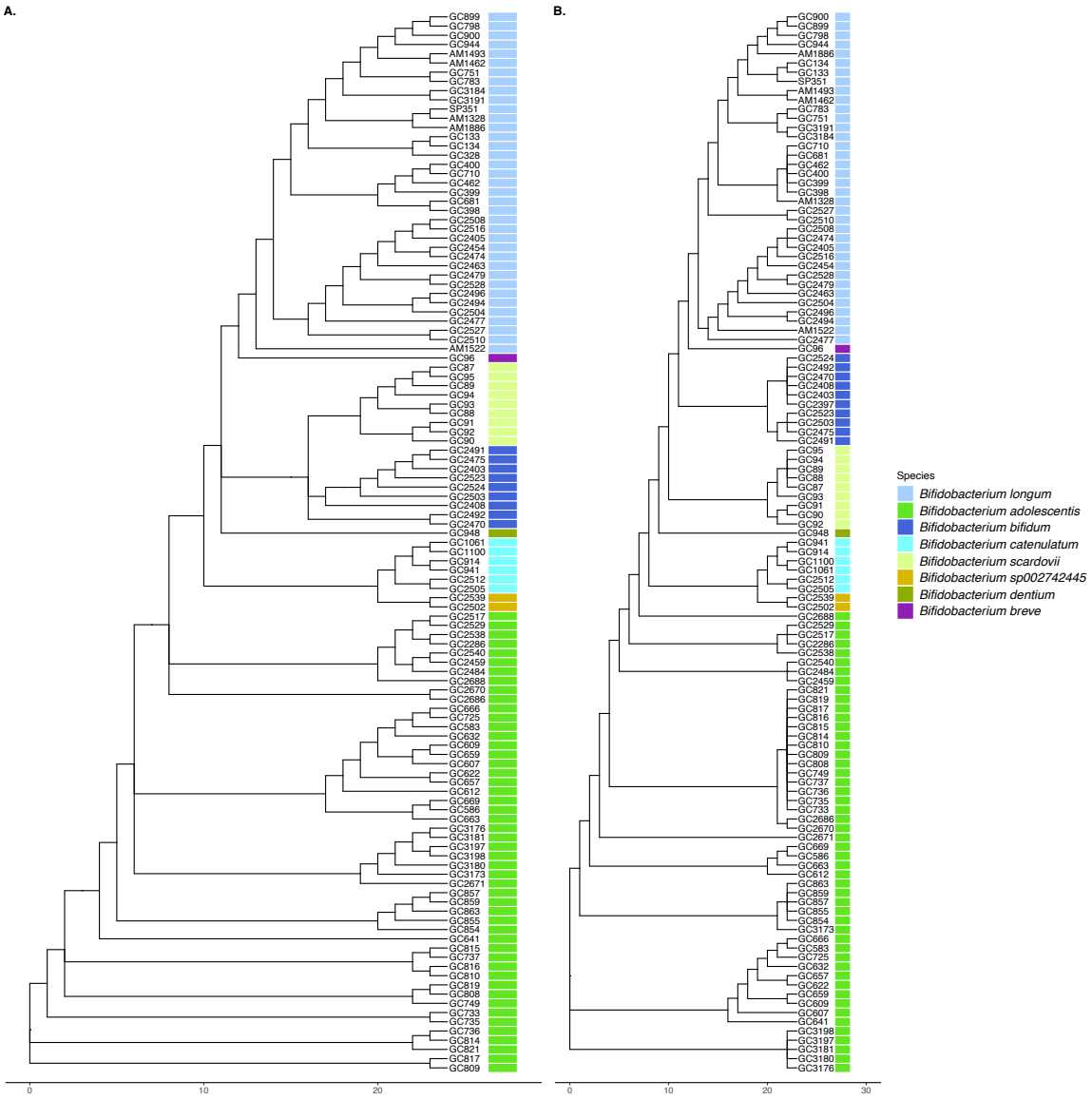


Figure 2.11. Genus-level phylogenetic trees of all *Bifidobacterium* species used in the study inferred using GTDB-Tk and Panaroo. A. A phylogenetic tree based on the core genome alignment of all Bifidobacterium species using Panaroo v.1.4.2 (Tonkin-Hill et al., 2020). **B.** A phylogenetic tree constructed based on the multiple sequence alignment of the 120 GTDB-Tk bacterial marker genes (bac120) present in strains (Chaumeil et al., 2020). FastTree v.2.1 was used to construct both trees (Price et al., 2010).

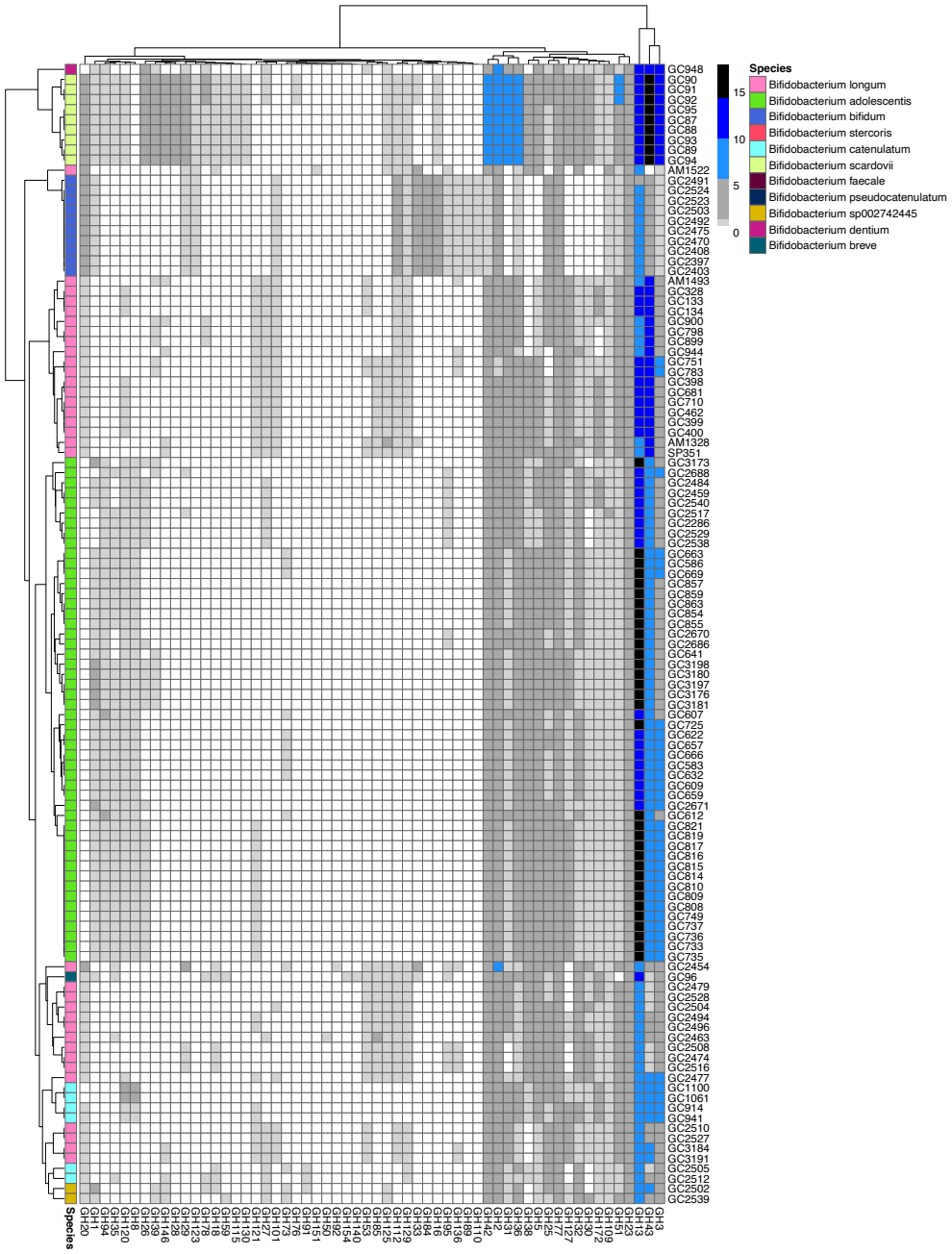
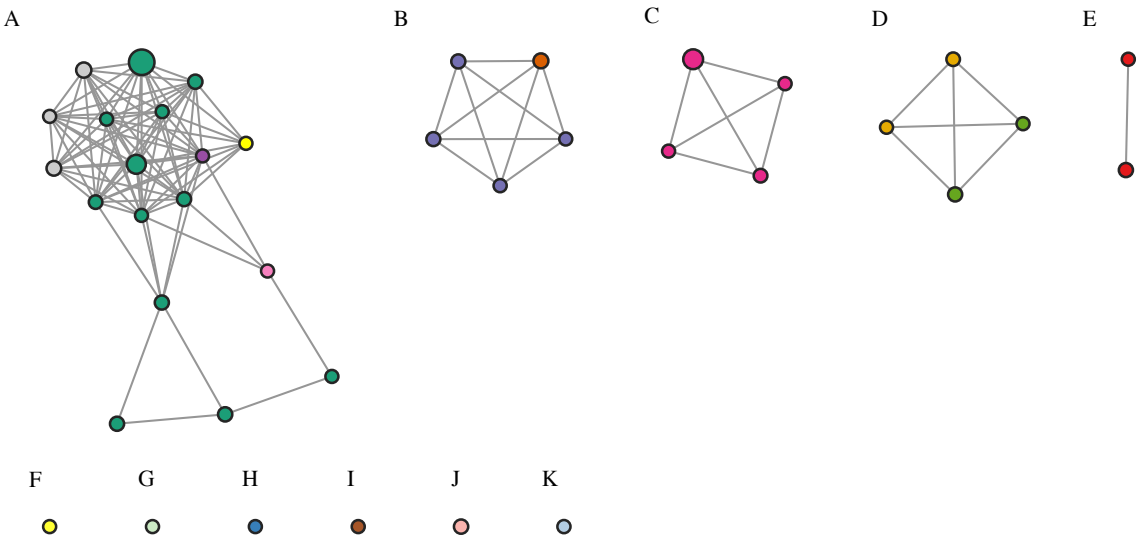


Figure 2.12. Clustering of bifidobacterial strains based on the abundance of GH genes. The heatmap illustrates the distribution and abundance of GH genes in bifidobacterial strains. White represents the absence of GH genes, bright grey indicates the presence of 1 GH gene, dark grey indicates 2-7 GH genes, light blue indicates 8-13 GH genes, dark blue indicates 14-19 GH genes, and black indicates 20-24 GH genes.

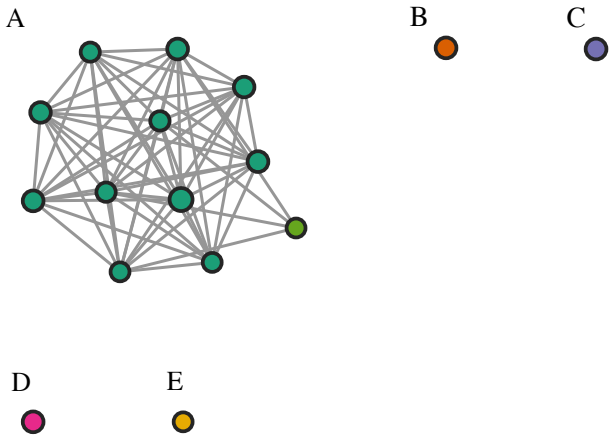


Activities in GH2	EC #
α-L-arabinopyranosidase	EC 3.2.1.-
β-1,3-galactosidase	EC 3.2.1.-
β-galacturonidase	EC 3.2.1.-
β-galacturonidase RGII specific	EC 3.2.1.-
Lactase	EC 3.2.1.108
Glycyrrhizin β-glucuronidase	EC 3.2.1.128
β-D-galactofuranosidase	EC 3.2.1.146
Mannosylglycoprotein endo-β-mannosidase	EC 3.2.1.152
Exo-β-1,4-glucosaminidase	EC 3.2.1.165
β-glucosidase	EC 3.2.1.21
β-galactosidase	EC 3.2.1.23
β-mannosidase	EC 3.2.1.25
β-glucuronidase	EC 3.2.1.31
Xylan β-1,4-xylosidase	EC 3.2.1.37
β-N-acetylhexosaminidase	EC 3.2.1.52
α-L-arabinofuranosidase	EC 3.2.1.55
β-galactosidase	EC 3.2.1.23
β-mannosidase	EC 3.2.1.25
β-glucuronidase	EC 3.2.1.31
Xylan β-1,4-xylosidase	EC 3.2.1.37
β-N-acetylhexosaminidase	EC 3.2.1.52
α-L-arabinofuranosidase	EC 3.2.1.55

Cluster	Sequence length	# of genes
$GH2_{A1}$	745	1
$GH2_{A2}$	197	31
$GH2_{A3}$	495	1
$GH2_{A4}$	336	1
$GH2_{A5}$	1045	191
$GH2_{B1}$	897	15
$GH2_{B2}$	911	20
$GH2_{C}$	720	59
$GH2_{D1}$	853	10
$GH2_{D2}$	800	6
$GH2_{E}$	639	10
$GH2_{F}$	591	1
$GH2_{G}$	607	1
$GH2_{H}$	1138	1
$GH2_{I}$	661	1
$GH2_{J}$	1936	10
$GH2_{K}$	948	2

% identity	# of nodes	# of edges
Full	361	27616
95	43	258
90	42	236
80	40	195
55	**39**	**190**

Figure 2.13. SSN of GH2 CAZy family. GH2 protein sequences were retrieved from 67 *Bifidobacterium* strains using dbCAN2 and submitted to EFI-EST to build the SSN (Zallot et al., 2019; Zhang et al., 2018). Cytoscape was used for visualization (Shannon et al., 2003). Each of the nodes contains sequences with more than 55% identity. Nodes are connected by an edge when the pairwise sequence identity is over 35%. A total of 39 nodes and 190 edges are present. The node size represents the number of genes present in the node. Alphabetical labels were used to represent different clusters of SSNs (e.g. $GH2_{A}$), and subscripts indicating numbers were used to show variations in protein sequence length and domain structure within the same clusters (e.g. $GH2_{A1}$).

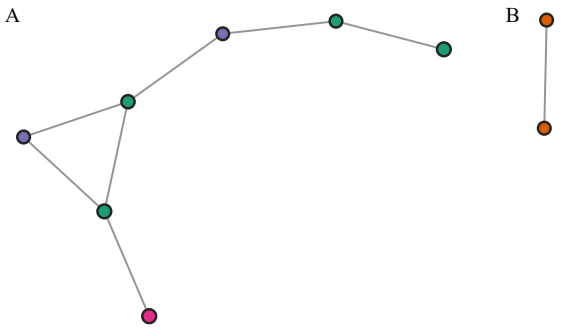


Cluster	Sequence length	# of genes
$GH20_{A1}$	697	79
$GH20_{A2}$	201	1
$GH20_{B}$	1061	10
$GH20_{C}$	1114	10
$GH20_{D}$	1628	10
$GH20_{E}$	670	1

Activities in GH20	EC #
β-1,6-N-acetylglucosaminidase	EC 3.2.1.-
β-N-acetyl-6-sulfo-glucosaminidase	EC 3.2.1.-
Lacto-N-biosidase	EC 3.2.1.140
β-N-acetylhexosaminidase	EC 3.2.1.52
Mannosyl-glycoprotein endo-β-N-acetylglucosaminidase	EC 3.2.1.96

% identity	# of nodes	# of edges
Full	111	3424
100	28	261
95	18	85
90	17	72
45	**16**	**61**

Figure 2.14. SSN of GH20 CAZy family. GH20 protein sequences were retrieved from 67 *Bifidobacterium* strains using dbCAN2 and submitted to EFI-EST to build the SSN (Zallot et al., 2019; Zhang et al., 2018). Cytoscape was used for visualization (Shannon et al., 2003). Each of the nodes contains sequences with more than 45% identity. Nodes are connected by an edge when the pairwise sequence identity is over 35%. A total of 16 nodes and 61 edges are present. The node size represents the number of genes present in the node. Alphabetical labels were used to represent different clusters of SSNs (e.g. $GH20_{A}$), and subscripts indicating numbers were used to show variations in protein sequence length and domain structure within the same clusters (e.g. $GH20_{A1}$).

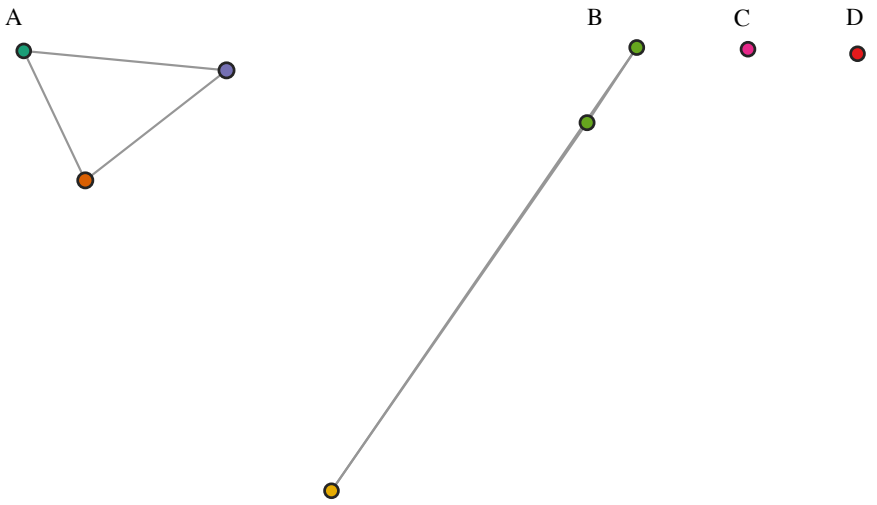


Cluster	Sequence length	# of genes
$GH29_{A1}$	459	24
$GH29_{A2}$	700	2
$GH29_{A3}$	1494	10
$GH29_{B}$	496	2

% identity	# of nodes	# of edges
Full	38	651
100	12	52
95	10	33
85	**9**	**26**

Activities in GH29	EC #
α-1,3-L-galactosidase	EC 3.2.1.-
α-1,4-L-fucosidase	EC 3.2.1.-
α-L-galactosidase	EC 3.2.1.-
α-L-glucosidase	EC 3.2.1.-
α-1,3-L-fucosidase	EC 3.2.1.111
α-1,6-L-fucosidase	EC 3.2.1.127
α-L-fucosidase	EC 3.2.1.51
α-1,2-L-fucosidase	EC 3.2.1.63

Figure 2.15. SSN of GH29 CAZy family. GH29 protein sequences were retrieved from 67 *Bifidobacterium* strains using dbCAN2 and submitted to EFI-EST to build the SSN (Zallot et al., 2019; Zhang et al., 2018). Cytoscape was used for visualization (Shannon et al., 2003). Each of the nodes contains sequences with more than 85% identity. Nodes are connected by an edge when the pairwise sequence identity is over 35%. A total of 9 nodes and 26 edges are present. The node size represents the number of genes present in the node. Alphabetical labels were used to represent different clusters of SSNs (e.g. $GH29_{A}$), and subscripts indicating numbers were used to show variations in protein sequence length and domain structure within the same clusters (e.g. $GH29_{A1}$).

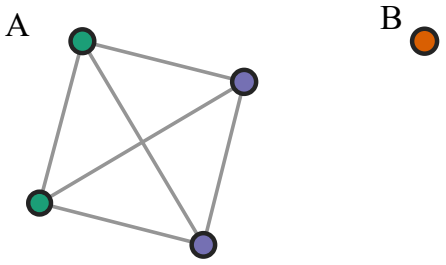


Cluster	Sequence length	# of genes
$GH33_{A1}$	395	1
$GH33_{A2}$	1796	10
$GH33_{A3}$	835	10
$GH33_{B1}$	765	5
$GH33_{B2}$	761	1
$GH33_{C}$	540	1
$GH33_{D}$	528	1

Activities in GH33	EC #
Trans-sialidase	EC 2.4.1.-
2-keto-3-deoxynononic acid hydrolase / KDNase	EC 3.2.1.-
Kdo hydrolase	EC 3.2.1.124
Exo-α-sialidase	EC 3.2.1.18
Anhydrosialidase	EC 4.2.2.15

% identity	# of nodes	# of edges
Full	29	373
100	9	25
95	**8**	**19**

Figure 2.16. SSN of GH33 CAZy family. GH33 protein sequences were retrieved from 67 *Bifidobacterium* strains using dbCAN2 and submitted to EFI-EST to build the SSN (Zallot et al., 2019; Zhang et al., 2018). Cytoscape was used for visualization (Shannon et al., 2003). Each of the nodes contains sequences with more than 95% identity. Nodes are connected by an edge when the pairwise sequence identity is over 35%. A total of 8 nodes and 19 edges are present. The node size represents the number of genes present in the node. Alphabetical labels were used to represent different clusters of SSNs (e.g. $GH33_{A}$), and subscripts indicating numbers were used to show variations in protein sequence length and domain structure within the same clusters (e.g. $GH33_{A1}$).



Cluster	Sequence length	# of genes
$GH95_{A1}$	784	6
$GH95_{A2}$	331	10
$GH95_{B}$	1960	10

Activities in GH95	EC#
α-L-galactosidase	EC 3.2.1.-
α-L-fucosidase	EC 3.2.1.51
α-1,2-L-fucosidase	EC 3.2.1.63

% identity	# of nodes	# of edges
Full	26	285
100	7	20
95	**5**	**9**

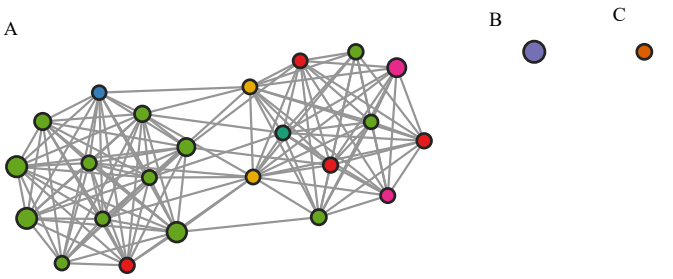
Figure 2.17. SSN of GH95 CAZy family. GH95 protein sequences were retrieved from 67 *Bifidobacterium* strains using dbCAN2 and submitted to EFI-EST to build the SSN (Zallot et al., 2019; Zhang et al., 2018). Cytoscape was used for visualization (Shannon et al., 2003). Each of the nodes contains sequences with more than 95% identity. Nodes are connected by an edge when the pairwise sequence identity is over 35%. A total of 5 nodes and 9 edges are present. The node size represents the number of genes present in the node. Alphabetical labels were used to represent different clusters of SSNs (e.g. $GH95_{A}$), and subscripts indicating numbers were used to show variations in protein sequence length and domain structure within the same clusters (e.g. $GH95_{A1}$).

Cluster	Sequence length	# of genes
$GH112_{A1}$	752	63
$GH112_{A2}$	185	1

Activities in GH112	EC#
β-galactoside phosphorylase	EC 2.4.1.-
β-1,3-galactosyl-N-acetylhexosamine phosphorylase	EC 2.4.1.211
D-galactosyl-β-1,4-L-rhamnose phosphorylase	EC 2.4.1.247

% identity	# of nodes	# of edges
Full	64	2016
100	17	136
95	6	15
85	4	6
80	**3**	**3**

Figure 2.18. SSN of GH112 CAZy family. GH112 protein sequences were retrieved from 67 *Bifidobacterium* strains using dbCAN2 and submitted to EFI-EST to build the SSN (Zallot et al., 2019; Zhang et al., 2018). Cytoscape was used for visualization (Shannon et al., 2003). Each of the nodes contains sequences with more than 80% identity. Nodes are connected by an edge when the pairwise sequence identity is over 35%. A total of 3 nodes and 3 edges are present. The node size represents the number of genes present in the node. Alphabetical labels were used to represent different clusters of SSNs (e.g. $GH112_{A}$), and subscripts indicating numbers were used to show variations in protein sequence length and domain structure within the same clusters (e.g. $GH112_{A1}$).

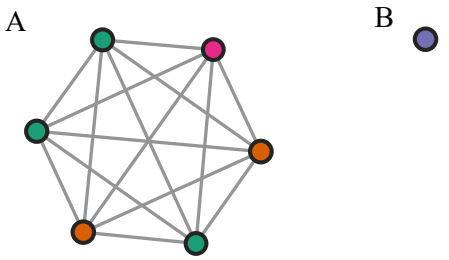


Cluster	Sequence length	# of genes
$GH42_{A1}$	689	216
$GH42_{A2}$	620	6
$GH42_{A3}$	478	34
$GH42_{A4}$	724	21
$GH42_{A5}$	873	1
$GH42_{A6}$	1236	3
$GH42_{B}$	705	9
$GH42_{C}$	709	51

Activities in GH42	EC #
α-L-arabinopyranosidase	EC 3.2.1.-
β-1,3-galactosidase	EC 3.2.1.-
β-galactosidase	EC 3.2.1.23
β-L-arabinosidase	EC 3.2.1.88

% identity	# of nodes	# of edges
Full	341	32533
100	82	2116
70	27	261
45	**25**	**217**

Figure 2.19. SSN of GH42 CAZy family. GH42 protein sequences were retrieved from 67 *Bifidobacterium* strains using dbCAN2 and submitted to EFI-EST to build the SSN (Zallot et al., 2019; Zhang et al., 2018). Cytoscape was used for visualization (Shannon et al., 2003). Each of the nodes contains sequences with more than 45% identity. Nodes are connected by an edge when the pairwise sequence identity is over 35%. A total of 25 nodes and 217 edges are present. The node size represents the number of genes present in the node. Alphabetical labels were used to represent different clusters of SSNs (e.g. $GH42_{A}$), and subscripts indicating numbers were used to show variations in protein sequence length and domain structure within the same clusters (e.g. $GH42_{A1}$).



Cluster	Sequence length	# of genes
$GH136_{A1}$	1706	1
$GH136_{A2}$	1368	9
$GH136_{A3}$	1612	6
$GH136_{B}$	975	4

Activities in GH136	EC#
Lacto-N-biosidase (Lewis antigen a/b specificity)	EC 3.2.1.-
Lacto-N-biosidase	EC 3.2.1.140

% identity	# of nodes	# of edges
Full	20	126
100	**7**	**15**

Figure 2.20. SSN of GH136 CAZy family. GH136 protein sequences were retrieved from 67 *Bifidobacterium* strains using dbCAN2 and submitted to EFI-EST to build the SSN (Zallot et al., 2019; Zhang et al., 2018). Cytoscape was used for visualization (Shannon et al., 2003). Each of the nodes contains sequences with more than 100% identity. Nodes are connected by an edge when the pairwise sequence identity is over 35%. A total of 7 nodes and 15 edges are present. The node size represents the number of genes present in the node. Alphabetical labels were used to represent different clusters of SSNs (e.g. $GH136_{A}$), and subscripts indicating numbers were used to show variations in protein sequence length and domain structure within the same clusters (e.g. $GH136_{A1}$).

Table 2.1. Summary of bifidobacterial strains in the lab collection. The sources of these strains are stool or sputum. The strain collection includes 8 *Bifidobacterium* species and 118 strains.

Species	**Number of strains**
Bifidobacterium adolescentis	50
Bifidobacterium longum	37
Bifidobacterium bifidum	10
Bifidobacterium scardovii	9
Bifidobacterium catenulatum	7
Bifidobacterium sp002742445	3
Bifidobacterium breve	1
Bifidobacterium dentium	1

Table 2.2. Summary of selected Bifidobacterium strains for HMO utilization assay[1].

Species	Strain	Species	Strain
Bifidobacterium adolescentis	GC2517	*Bifidobacterium faecale*	GC583
	GC3181		GC586
	GC2671		GC607
	GC2459	*Bifidobacterium pseudocatenulatum*	GC941
	GC2688		GC1061
	GC2529		GC914
	GC2686	*Bifidobacterium scardovii*	GC90
	GC3173		GC94
	GC641		GC95
	GC749		GC87
	GC821	*Bifidobacterium sp002742445*	GC2539
	GC612		GC2464
Bifidobacterium bifidum	GC2403		GC2502
	GC2524	*Bifidobacterium breve*	GC96
	GC2491	*Bifidobacterium catenulatum*	GC2512
	GC2503		GC1099
	GC2475		GC2505
	GC2470		GC1100
	GC2397	*Bifidobacterium stercorsis*	GC859
	GC2408		GC666
	GC2492		GC659
	GC2523		GC854
Bifidobacterium longum	GC783		GC657
	SP351		GC863
	GC134		GC659
	GC2527		GC663
	GC2454		GC725
	GC900	*Fusicatenibacter saccharivorans*	GC474
	GC400		GC313
	GC2463	*Coprococcus eutactus*	GC567
	GC2477	*Enterococcus faecium*	GC33
	GC3184	*Blautia luti*	GC555
	AM1493	*Blautia producta*	GC553
	AM1328	*Lactobacillus rhamnosus*	GC38
	AM1522	*Bacteroides cellusilyticus*	GC234
	AM1321	*Bacteroides ovatus*	GC401
	GC2504		GC137
	GC2528	*Bacteroides xylanisolvens*	GC232
	GC2474		
	GC681		

[1]Bifibacterial strains were selected based on phenotypic screens and comparative genomics.

Table 2.3. The distribution of GH family within the collection of 67 *Bifidobacterium* genomes, with 58 GH families and a total of 7296 GHs identified.

GH family	Number of GH
GH1	76
GH101	34
GH109	117
GH110	10
GH112	64
GH115	9
GH120	75
GH121	41
GH123	20
GH125	36
GH127	146
GH129	56
GH13	1375
GH130	9
GH136	20
GH140	1
GH146	34
GH151	1
GH154	1
GH16	29
GH172	120
GH18	18
GH2	361
GH20	111
GH23	324
GH25	214
GH26	49
GH27	38
GH28	18

GH family	Number of GH
GH29	38
GH3	588
GH30	116
GH31	199
GH32	167
GH33	29
GH35	56
GH36	287
GH38	203
GH39	35
GH42	341
GH43	870
GH5	206
GH50	1
GH51	250
GH53	25
GH59	11
GH73	14
GH76	3
GH77	230
GH78	19
GH8	56
GH84	20
GH85	23
GH89	10
GH91	3
GH92	1
GH94	62
GH95	26

CHAPTER 3: Mapping Glycoside Hydrolases to HMO Utilization with Machine Learning and to Metagenomic Data from CHILD[1]

3.1 Introduction

In Chapter 2, *Bifidobacterium* strain differences in families were identified and GH family subtypes were classified using SSN. I hypothesized that these GH subtypes would act on specific HMOs. Using comparative genomics and functional assays, I prioritized a set of *Bifidobacterium* isolates that represented species and strain diversity in our collection (**Table 2.2**) for further analysis. By combining HMO utilization data with GH subtype distribution in strains, I sought to assign some GH subtypes to the utilization of specific HMOs. Furthermore, using information from strain HMO utilization and GH subtypes, I can revisit the CHILD metagenomic data to look for gene-specific enrichment in asthma phenotypes.

To investigate how specific GH enzymes are involved in HMO metabolism, HMO utilization of strains prioritized in Chapter 2 was carried out. The study

[1] Data presented in this work was facilitated through collaborations. Dr. Lars Bode provided the purified pooled HMOs for my growth experiments, and his group carried out the analysis of HMO degradation. Dr. Nick Dimonaco assisted with the implementation of the WEKA decision tree software. Dr. Shahrokh Shekarriz carried out the metagenomic read mapping and assisted with the analysis. Dr. Charisse Peterson provided the CHILD metagenomic data.

aimed to evaluate the HMO utilization capability of the selected strains and determine the specific HMO metabolism pathway for bifidobacterial strains by gathering data from the SSN subtypes from Chapter 2 and the HMO utilization assay. HMO degradation reaction was assigned to GH subtype genes using decision trees. A decision tree is a machine-learning technique that forecasts outcomes based on input data. It comprises nodes and branches to illustrate how decisions are made under various conditions (Mitchell, 1997). In this chapter, a machine learning decision tree algorithm applied to HMO degradation and GH subtypes (from the SSN analysis) was used to identify the specific GH genes linked to high HMO degradation.

The metagenomic data from the CHILD study was mapped against the strains and their GH subtypes to investigate whether a stronger association of GH subtypes with asthma protection is observed compared to *B. infantis*. Two strategies were employed for metagenomic read mapping. First, nine bifidobacteria strains, including those that efficiently degraded HMOs and strains that did not, were used for mapping the metagenomic reads. Second, the GH subtypes were used to create a database. Mapping the metagenomic reads against this database allowed for the identification of genes associated with asthma protection. This unbiased approach aimed to identify GH subtype genes enriched in individuals who developed asthma versus those who did not.

3.2 Methods

3.2.1 Growth curve generation and HMO utilization assay

Selected strains were cultured from the frozen stocks into the BHI3 agar and incubated for 24 to 48 hours in the ANOXOMAT jar (10% H_2, 10% CO_2, 85% N_2). This includes specific bifidobacteria isolates prioritized in Chapter 2 (**Table 2.2**) and a few non-bifidobacteria strains for comparison (**Table 3.1**). A single colony of each strain was picked and grown overnight in 1 mL of PYG broth in an anaerobic chamber. The PYG broth, which was supplemented with 15 g/L of pooled HMOs (pHMOs, a gift from Dr. Lars Bode, University of San Diego, California (UCSD), USA) was inoculated with 5% (v/v) seed culture and overlaid with 50 μL of sterile mineral oil to prevent evaporation. The 96-well plate was transferred to an Epoch 2 microplate spectrophotometer under anaerobic conditions and incubated for 48 hours at 37℃ , while OD_{600} readings were recorded every 30 minutes, with each reading used to generate a growth curve. After 6 h, 12 h, and 24 h of incubation, 25 μL of spent media was collected for glycoprofiling (**Figure 3.1**). The spent media was centrifuged at 4,000 rpm for 5 minutes and stored at -80℃.

3.2.2 Glycoprofiling of HMO

Glycoprofiling was done by Dr. Lars Bode from UCSD (University of San Diego, California). HMOs were analyzed using high-performance fluorescence liquid chromatography (HPLC) after labeling them with fluorescent tag 2-

aminobenamide (2AB), as previously described (Autran et al., 2018; Bode et al., 2012). Nineteen individual HMOs structures were quantified, based on the retention time and mass spectrometry (**Figure 1.1**): 2'-fucosyllactose (2'FL), 3-fucosyllactose (3-FL), 3'-sialyllactose (3'-SL), 6'-sialyllactose (6'SL), difucosyllactose (DFLac), difucosyllacto-N-hexaose (DFLNH), difucosyllacto-N-tetraose (DFLNT), disialyllacto-N-hexaose (DSLNH), disialyllacto-N-tetraose (DSLNT), fucodisialyllacto-N-hexaose (FDSLNH), fucosyllacto-N-hexaose (FLNH), lacto-N-fucopentaose (LNFP), lacto-N-hexaose (LNH), lacto-N-neotetraose (LNnT), lacto-N-tetraose (LNT), sialyl-lacto-N-tetraose b (LSTb), and sialyl-lacto-N-tetraose c (LSTc).

The HMO glycoprofiling data from UCSD included the percentage of each of the nineteen HMOs remaining compared to the standard controls at each time point 6, 12, and 24h. Using the percentages of HMOs remaining at the 24h timepoint, degradation values were categorized with a threshold of 80. Percentages of HMOs lower than 80, indicating HMO degradation, were converted to 1; values between 80 and 120, indicating no degradation, were converted to 0; and values higher than 120, indicating HMO accumulation, were converted to 2. These values were used to visualize the heatmap of HMO utilization assay data.

3.2.3 Machine Learning using WEKA

For machine learning analysis, strains were excluded from the study when their OD_{600} readings were smaller than 0.12, indicating no growth, and when no

HMO degradation was observed, but instead accumulated. For specific HMOs, the data was simplified as follows: values below 80% were converted to 1, indicating degradation, while 80% or higher values were converted to 0, indicating no degradation or potential accumulation. We then identified GH families associated with HMO degradation and gathered their corresponding GH subtypes. The HMO degradation data and GH subtypes were compiled and saved as CSV files. These CSV files were subsequently converted into WEKA's Attribute-Relation File Format (ARFF) using a custom Python script. For each HMO, we utilized the J48 algorithm as a classifier to analyze the data and generate decision trees with a minimum number of two objects, covering all nineteen HMOs.

3.2.4 Metagenomics study design

A total of 76 metagenomic read data from the CHILD study was provided by Dr. Charisse Peterson (University of British Columbia) (**Table 3.3**). Two different approaches were employed for metagenomics read mapping (**Figure 3.2**). First, metagenomics read mapping was performed against nine strains comprising of good and poor HMO degraders selected based on the HMO degradation data. Second, metagenomics reads were mapped against all GH subtype genes. A reference database was created by extracting the gene sequences for all strains' GH subtypes and CD-HIT was performed at 99% (Fu et al., 2012). The read mapping was performed using bwa-mem, and the resulting SAM file was converted to a BAM file using SAM tools (Li et al., 2009; Li &

Durbin, 2009). HTSeq was used to count the number of aligned reads (Anders et al., 2015). Further analysis, including normalization, was done using a generalized linear model in R using the following packages: glmmTMB v.1.1.9, ggplot2 v.3.5.1, tidyverse v.2.0.0, dplyr v.1.1.4 and ggpubr v.0.6.0 (Brooks et al., 2017; Kassambara, 2020; Wickham, 2016; Averick, et al., 2019; François, et al., 2019).

3.3 Results

3.3.1 *Bifidobacterium* strain-specific HMO degradation profiles

To assess the capability of *Bifidobacterium* species to metabolize HMOs, HMO utilization assay and glycoprofiling were conducted on 78 strains. This group included *Bifidobacterium* species as well as other species known to metabolize HMOs, such as *Bacteroides, Fusicatenibacter, Coprococcus, Enterococcus, Lactobacillus,* and *Blautia* strains (Marcobal et al., 2011; Ward et al., 2006). The concentrations of HMOs remaining after bacterial culture compared to the standard controls were quantified at 6 h, 12 h, and 24 h time points. The heatmap and hierarchical clustering of the growth data are displayed in **Figure 3.3**. The heatmap of the HMO degradation pattern at a 24 h time point displays the strain heterogeneity of HMO degradation. The growth of each strain is confirmed by the OD_{600} readings obtained during the growth curve generation. This data provides insight into which HMO is utilized first and which is not utilized for each strain.

In general, *B. longum* and *B. bifidum* metabolized HMOs well compared to other species. Some *B. longum* strains degrade HMOs better than other strains and accumulation of HMOs was observed. Three *B. bifidum* strains (GC2470, GC2475 and GC2492) could utilize all nineteen HMO structures. However, not all *B. longum* and *B. bifidum* strains utilized HMOs well, indicating both strain and species heterogeneity. The *B. infantis* strain, AM1522, efficiently utilized most of the HMO structures, aligning with previous literature that *B. infantis* is a good HMO degrader due to its dedicated HMO gene cluster (Chichlowski et al., 2020; Underwood et al., 2014; Ward et al., 2006). Despite 2'FL being one of the most common HMOs to assess HMO degradation activity, it was utilized by 17 strains out of 78 strains (Hegar et al., 2019). Three *B. bifidum* strains could degrade all 19 HMOs, and the *B. infantis* strain, *AM1522*, grew well in pHMOs-supplemented media and degraded most HMO structures. LNT was the most commonly degraded HMO, and all *B. longum* strains were able to degrade LNT. HMO structures were accumulated across several *Bifidobacterium* species, *B. adolescentis, B. pseudocatenulatum, B. catenulatum, B. scardovii,* and *B. stercorsis*. These were HMOs that could be generated from larger HMOs or glycoside transferase activity and these accumulations were not further investigated.

3.3.2 Using machine learning to identify GH subtypes associated with HMO degradation and assigning enzymatic reactions

A decision tree was visualized for each of the nineteen HMOs to predict the GH subtypes associated with HMO degradation. The degradation threshold was determined as 80% and a few strains with low growth (OD_{600} < 0.12) and no HMO utilization were excluded from the analysis as they could represent false negatives and adversely affect the decision tree analysis. The number of nodes and branches present in the decision tree differed for each HMO. In the decision tree for 2'FL (**Figure 3.4.A**), the branch corresponding to $GH20_B$ = 1, which indicates the presence of $GH20_B$, comprises 8 of 52 cases in the data. The numbers in the parentheses represent how many cases fall into the category, with the number after the slash symbol in the parentheses representing the misclassified cases. As displayed in Figure 3.4.A, the decision tree predicts that the absence of $GH20_B$ ($GH20_B$ = 0) and $GH136_B$ leads to no degradation (0) of 2'FL in 36 of 52 instances, excluding the three misclassified cases. The absence of $GH20_B$ ($GH20_B$ = 0) but the presence of $GH136_B$ ($GH136_B$ = 1) leads to degradation (1) of 2'FL in 2 of 52 cases.

The decision tree predicted that $GH20_B$ and $GH29_{A1}$ are associated with the degradation of 3-FL (**Figure 3.4**.**B**). GH20 comprises β-N-acetylglucosaminidase and lacto-N-biosidase, which cleave the linkage between galactose and N-acetyl-D-glucosamine (Kitaoka, 2012; Lombard et al., 2014). However, N-acetyl-D-glucosamine is not a building block of 3-FL. GH29 includes

α-galactosidase, α-glucosidase and α -fucosidase (Lombard et al., 2014). Among these, 1,3-α-L-fucosidase hydrolyzes the fucose unit from galactose, making $GH29_{A1}$ responsible for the hydrolysis of fucose in 3-FL (Ashida et al., 2009). Similarly, $GH20_B$ and $GH136_B$ were predicted to be involved in DFLNH degradation (**Figure 3.5.B**). Both GH20 and GH136 contain lacto-N-biosidase. However, lacto-N-biosidase from GH20 is responsible for cleaving β-1,3 and β-1,6 linkage, whereas the enzyme from GH136 cleaves only the β-1,3 linkage (Lombard et al., 2014). Thus, the β-1,3 linkage between N-acetyl-D-glucosamine and galactose in DFLNH can be cleaved by either $GH20_B$ or $GH136_B$, while the β-1,6 linkage is cleaved exclusively by $GH20_B$.

The complexity of the decision tree varies. For instance, the decision tree for 3'SL has one decision node (**Figure 3.4.C**), while the decision tree for LNH has three decision nodes (**Figure 3.6.D**). Of the 19 decision trees, 13 had $GH20_B$ as the root node, which is the topmost feature of the tree representing the first decision (Mitchell, 1997).

3.3.3 Metagenomic read mapping to identify GH genes that are enriched and depleted in asthma samples from the CHILD study

Metagenomic read mapping was performed using two complementary approaches as previously described. Based on the HMO utilization assay data, six good and three poor HMO utilizing *Bifidobacterium* strains were selected (**Table 3.2**). These nine stains included two *Bifidobacterium* species, *B. bifidum* and *B. longum*. The relative abundance of these strains was calculated based on

the number of total reads present in each metagenomic sample. Nine strains were highly abundant in many of the metagenomic samples. AM1522 strain, *B. infantis*, was particularly abundant in some samples from individuals diagnosed with possible asthma at age 5 (**Figure 3.9.A**). Compared to samples collected at 3 months, there was generally a decrease in the abundance of *B. bifidum* and *B. longum* by 1 year (**Figure 3.9.B**). Additionally, the relative abundance of the AM1522 strain decreased from 3 months to 1 year. The metagenomic samples were organized based on StrainPhlAn results, which characterize strains based on species-specific marker sequence differences (**Figure 3.9.C**) (Truong et al., 2017). “Infantis” indicates a high abundance of *B. infantis*, “Other” indicates the presence of other *B. longum* subspecies, and “None” indicates that no *B. longum* subspecies were detected. Most samples dominated by *B. infantis* according to StrainPhlAn had a high relative abundance of AM1522, except for one sample (7_118038_1_4).

The genome coverage was widely distributed in samples not diagnosed with asthma at age 5. Three samples—two from the non-asthma group and one from the possible asthma group—had high abundance and coverage of AM1522, suggesting this strain best represents these subjects (**Figure 3.10**). Metagenomic read mapping was also performed against GH subtypes identified using SSN (Chapter 2). **Figure 3.11** displays that $GH2_{K}$, $GH13_{A5}$, $GH13_{B1}$, $GH2_{I}$, $GH136_{A1}$, $GH2_{E}$, $GH13_{B5}$ and $GH95_{A2}$ are enriched in the asthma group compared to

individuals who did not develop asthma. Many GH subtypes were depleted in the asthma group, with $GH2_{A4}$ and $GH2_{A1}$ being particularly depleted.

3.4 Discussion

In Chapter 2, SSN was used to assign GH subtype genes which are hypothesized to have different substrate specificity and enzymatic activity. To assign enzymatic reactions and functions to GH subtype genes, *Bifidobacterium* species' ability to utilize HMOs was evaluated. A total of 76 selected strains were subjected to HMO utilization assay and glycoprofiling in collaboration with Dr. Lars Bode (UCSD). Glycoprofiling measured HMO concentrations at 3 different time points (6 h, 12 h and 24 h). For further analysis, the degradation profile at 24 h time point was used.

In **Figure 3.3**, "No degradation" indicates that no degradation of the HMO structure was observed. This could be possible that these *Bifidobacterium* strains may need more than 24 hours to metabolize HMOs, lack growth components in the media, or simply cannot utilize those HMO structures. "Accumulation" suggests that the HMO structure increased over time, possibly due to the release of new compounds that are not further degraded while degrading other HMOs. For example, LNFP III accumulation might occur as LNnT undergoes sialyation. Alternatively, LNT accumulation may result from the cleavage of N-acetyl-D-glucosamine in DSLNT releasing LNT. To clarify this, evaluating the growth of these *Bifidobacterium* strains on individual HMOs rather than pooled HMOs could provide a better understanding.

In Chapter 2, I observed strain-level differences in carbohydrate utilization, which were also evident in HMO utilization. *Bifidobacterium* species isolated from humans are generally categorized into two types, adult-type and infant-type *Bifidobacterium* (Lin et al., 2022). Adult-type *Bifidobacterium*, found in adult stool samples, includes *B. adolescentis, B. pseudocatenulatum,* and *B. catenulatum* (Turroni et al., 2011; Wong et al., 2018). In contrast, infant-type *Bifidobacterium* encompasses species like *B. infantis, B. longum, B. breve* and *B. bifidum* and they possess genes and enzymes specific for utilizing HMOs, giving them a growth advantage over other microbes (Asakuma et al., 2011). Our study observed that *B. longum* and *B. bifidum* strains efficiently degraded HMOs compared to other *Bifidobacterium* species. While 2'FL is a common prebiotic added to infant formulas and used to assess HMO degradation, many strains did not utilize it, whereas LNT was the most commonly degraded HMO (Puccio et al., 2017). Duar et al. (2020) assessed the ability of 12 *Bifidobacterium* strains to metabolize LNT, LNnT and 2'FL and concluded that these strains metabolized LNT and LNnT more effectively than 2'FL. This suggests LNT is a strong candidate for use as a prebiotic and aligns with earlier research that examined LNT as an additive of infant formula (Hu et al., 2023).

Machine learning was used to identify the GH subtype genes that are associated with HMO degradation. The model was trained on the HMO glycoprofiling data at 24 h timepoint. Among 19 decision trees, $GH20_B$ was most often found as the root node. GH20 enzymes are responsible for cleaving N-

acetyl-D-glucosamine containing substrates but with varying specificities (Lombard et al., 2014). $GH20_B$ was only present in *B. bifidum* strains in our collection and may be responsible for the strain's ability to degrade a wide variety of HMOs, as indicated by our glycoprofiling data. Although decision trees helped assign GH subtypes to enzymatic reactions, some trees lacked resolution. This may be due to inaccurate predictions of GH subtype activity. To assess this, the predicted GH subtypes could be cloned and tested with the substrate. Mass spectrometry analysis could determine if the substrate is degraded or not. Alternatively, predicted GH subtypes may be associated with other GH enzymes or carbohydrate-binding modules. For instance, for 3'SL, $GH20_B$ activity may be inhibited by the sialic acid but it may co-occur with other sialidases. Expanding the analysis to include all GH families and higher resolution mass spectrometry data where degradation products would be identified would resolve some of these issues. It would also allow more sophisticated analysis and improve the assignment of GH subtypes to specific reactions.

A previous metagenomic study has linked the protective effect of breastfeeding to the enrichment of *B. infantis* (Dai et al., 2023). However, I hypothesized that the protective effect is not solely due to the presence of *B. infantis*, but rather to a specific set of genes that may be found in some strains of *Bifidobacterium* species. To identify GH subtypes associated with asthma protection, metagenomic read mapping was performed with two approaches. Metagenomic reads mapped against nine selected strains revealed that the

AM1522 (*B. infantis*) strain was relatively abundant even in samples with asthma phenotype, suggesting that *B. infantis* colonization alone does not confer asthma protection. A decrease in the relative abundance of *B. bifidum* and *B. longum* was observed at 3 months compared to 1 year, coinciding with an increased diversity in the infant gut microbiota. As infants transition from milk feeding to the introduction of solid foods at 6 months, their gut microbiota undergoes changes and becomes diverse (Differding et al., 2020). During this period, alpha diversity increases, and the microbial community shifts from being dominated by *Bifidobacterium* to being dominated by *Bacteroidetes* and *Firmicutes*. The relative abundance of the AM1522 strain decreases over time as *B. infantis* is one of the most prevalent microbes in early life during breastfeeding.

Metagenomic read mapping was carried out against GH subtype genes to look for gene-specific enrichment in asthma phenotypes. $GH2_{A4}$ and $GH2_{A1}$ genes were highly depleted in asthma phenotypes and were only present in one *B. adolescentis* strain (GC641) in our collection, despite GC641 degrading only two of nineteen HMOs. These genes may be present in other *B. infantis* isolates, or other strains may contribute to the protection against asthma. This suggests that asthma development is not solely influenced by one *Bifidobacterium* strain. Rather, factors associated with risk likely vary among different phenotypes. Larger datasets could offer more precise predictions regarding the impact of *Bifidobacterium* strains and GH gene distributions on asthma risk, which will be the focus of future studies.

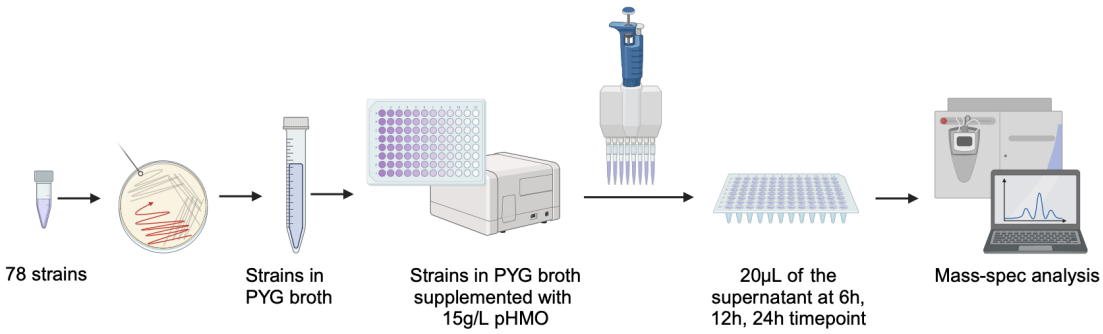


Figure 3.1. Workflow describing the growth curve experiment and HMO utilization assay. Supernatant of bacterial culture was sent to UCSD for glycoprofilling. Figure constructed using BioRender.

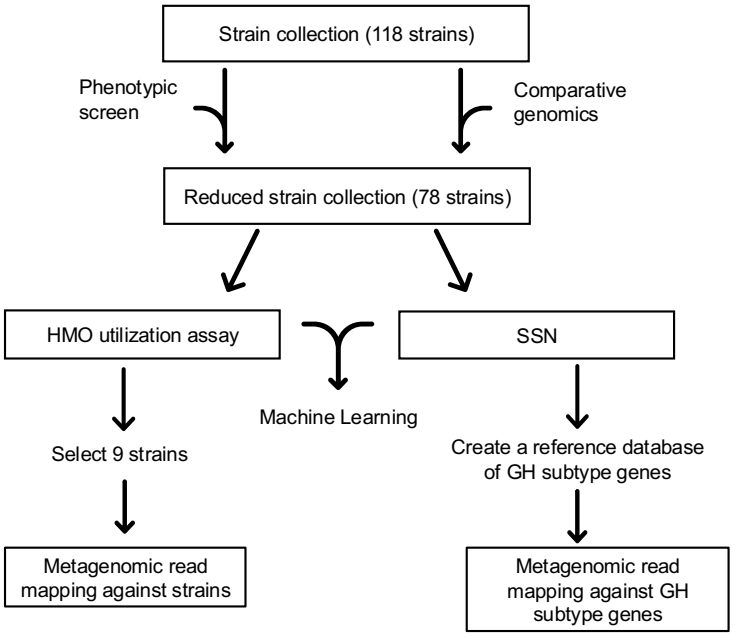


Figure 3.2. Workflow of metagenomic read mapping. Two strategies were used for metagenomic read mapping. First, metagenomic reads were mapped against nine selected strains. Second, metagenomic reads were mapped against GH subtype genes.

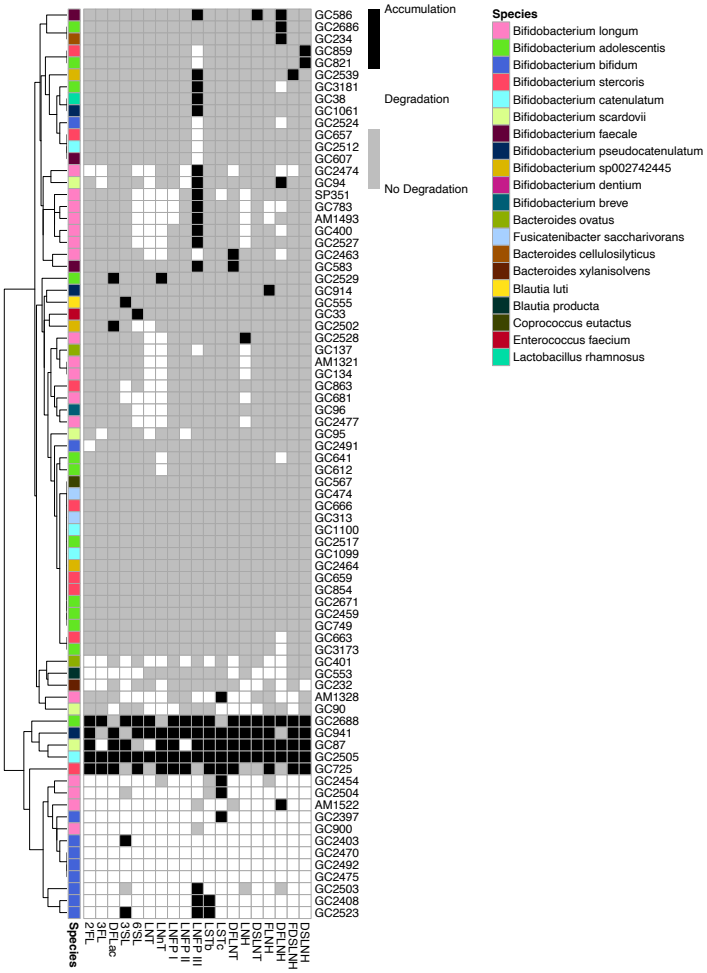
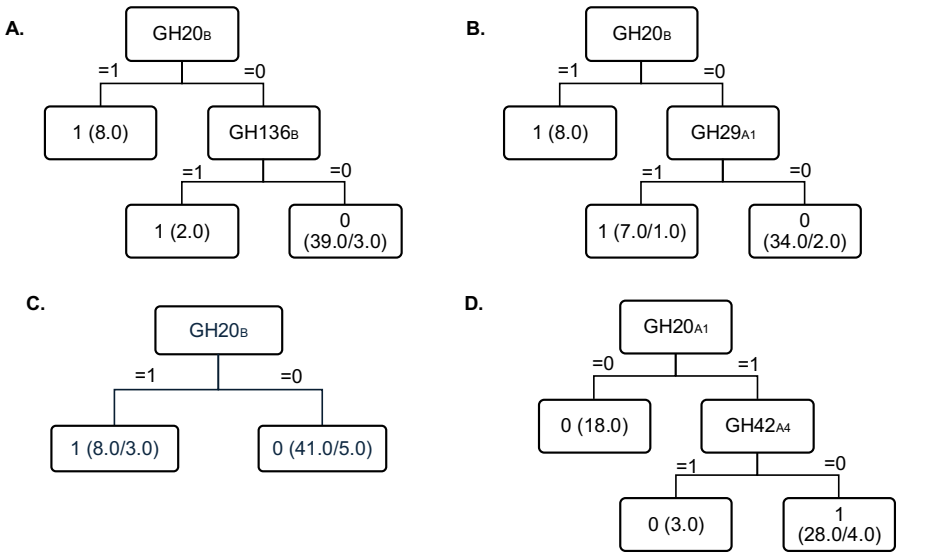


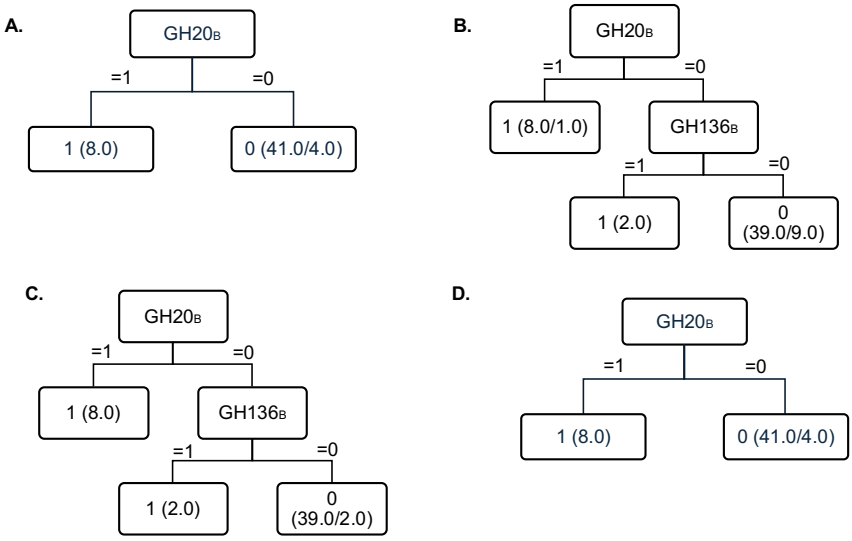
Figure 3.3. HMO utilization profiles of bifidobacterial strains at the 24 hr time point. Seventy-eight strains were cultured in media supplemented with 15 g/L pHMOs. Glycoprofiling was conducted at the 24 hr time point to track the utilization of each of the nineteen HMOs. Grey indicates no degradation (80-120), white indicates degradation (<80), and black indicates accumulation (>120).



Substrate	Predicted GH subtype	Predicted reaction / Alternative explanation
A. 2'FL	$GH20_B$ (O) → degradation (**8**) $GH20_B$ (X) & $GH136_B$ (O) → degradation (**2**)	GH20 and GH136 are not predicted to have this activity
B. 3-FL	$GH20_B$ (O) → degradation (**8**) $GH20_B$ (X) & $GH29_{A1}$ (O) → degradation (**8**)	← $GH29_{A1}$
C. 3'SL	$GH20_B$ (O) → degradation (**10**)	GH20 family is not predicted to have this activity
D. 6SL	$GH20_{A1}$ (O) & $GH42_{A4}$ (X) → degradation (**24**)	GH20 family is not predicted to have this activity

Glc
Gal
NeuAc
Fuc
GlcNAc

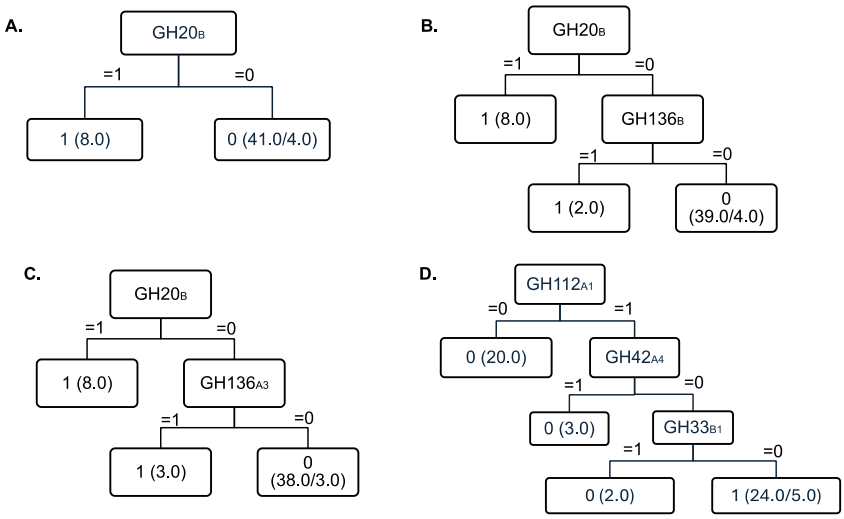
Figure 3.4. J48 decision tree predicting the GH subtypes highly associated with HMO degradation. A. 2'FL. **B.** 3-FL. **C.** 3'SL. **D.** 6'SL. The numbers in parentheses represent the total sample count and the instances that were classified incorrectly. The table displays information about the GH subtypes predicted to be involved in the enzymatic reaction of each substrate. Potential cleavage sites of the glycosidic linkage are indicated with an arrow.



Substrate	Predicted GH subtype	Predicted reaction / Alternative explanation
A. DFLac	$GH20_B$ (O) → degradation (**12**)	GH20 family is not predicted to have this activity
B. DFLNH	$GH20_B$ (O) → degradation (**7**) $GH20_B$ (X) & $GH136_B$ (O) → degradation (**11**)	$GH20_B$ $GH20_B$ and/or $GH136_B$
C. DFLNT	$GH20_B$ (O) → degradation (**8**) $GH20_B$ (X) & $GH136_B$ (O) → degradation (**4**)	GH20 and GH136 are not predicted to have this activity
D. DSLNT	$GH20_B$ (O) → degradation (**12**)	$GH20_B$

Glc
Gal
NeuAc
Fuc
GlcNAc

Figure 3.5. J48 decision tree predicting the GH subtypes highly associated with HMO degradation. A. DFLac. **B.** DFLNH. **C.** DFLNT. **D.** DSLNT. The numbers in parentheses represent the total sample count and the instances that were classified incorrectly. The table displays information about the GH subtypes predicted to be involved in the enzymatic reaction of each substrate.



Substrate	Predicted GH subtype	Predicted reaction / Alternative explanation
A. DSLNH	$GH20_B$ (O) → degradation (**12**)	$GH20_B$ can act on these linkages
B. FDSLNH I	$GH20_B$ (O) → degradation (**8**) $GH20_B$ (X) & $GH20_B$ (O) → degradation (**6**)	$GH20_B$ $GH20_B$ and/or $GH136_B$
C. FLNH	$GH20_B$ (O) → degradation (**8**) $GH20_B$ (X) & $GH136_{A3}$ (O) → degradation (**6**)	$GH20_B$ $GH20_B$ and/or $GH136_B$
D. LNH	$GH112_{A1}$ (O) & $GH42_{A4}$ (X) & $GH33_{B1}$ (X) → degradation (**19**)	GH112 family is not predicted to have this activity

Figure 3.6. J48 decision tree predicting the GH subtypes highly associated with HMO degradation. A. DSLNH. **B.** FDSLNH I. **C.** FLNH. **D.** LNH. The numbers in parentheses represent the total sample count and the instances that were classified incorrectly. The table displays information about the GH subtypes predicted to be involved in the enzymatic reaction of each substrate. Potential cleavage sites of the glycosidic linkage are indicated with an arrow.

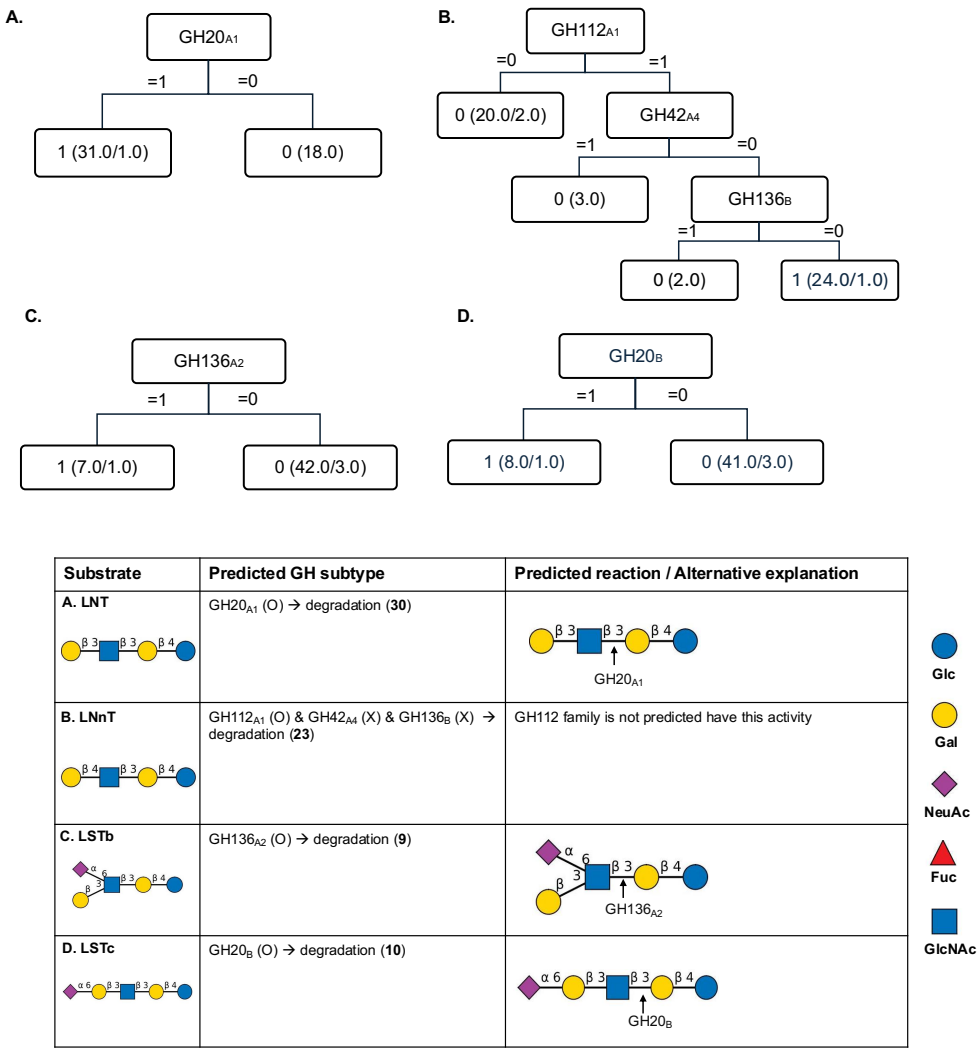
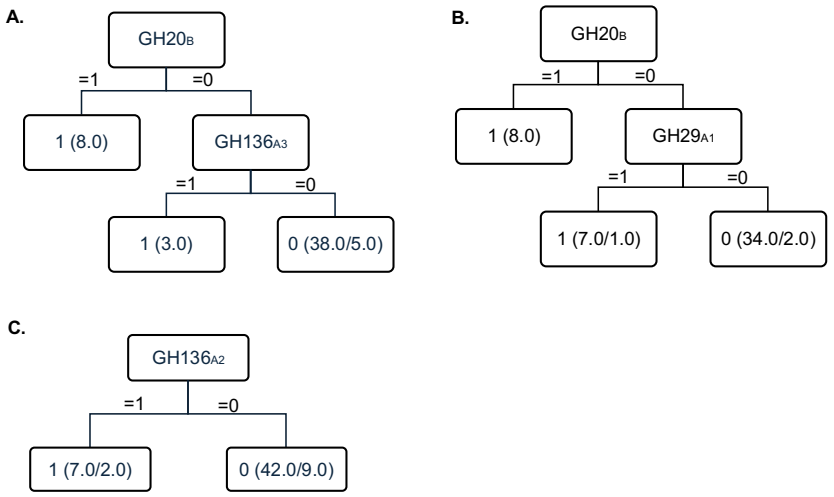


Figure 3.7. J48 decision tree predicting the GH subtypes highly associated with HMO degradation. A. LNT. **B.** LNnT. **C.** LSTb. **D.** LSTc. The numbers in parentheses represent the total sample count and the instances that were classified incorrectly. The table displays information about the GH subtypes predicted to be involved in the enzymatic reaction of each substrate. Potential cleavage sites of the glycosidic linkage are indicated with an arrow.



Substrate	Predicted GH subtype	Predicted reaction / Alternative explanation
A. LNFP I	$GH20_B$ (O) → degradation (**8**) $GH20_B$ (X) & $GH136_{A3}$ (O) → degradation (**8**)	$GH136_{A3}$ and/or $GH20_B$
B. LNFP II	$GH20_B$ (O) → degradation (**8**) $GH20_B$ (X) & $GH29_{A1}$ (O) → degradation (**8**)	$GH29_{A1}$ $GH20_B$
C. LNFP III	$GH136_{A2}$ (O) → degradation (**14**)	$GH136_{A2}$

Glc
Gal
NeuAc
Fuc
GlcNAc

Figure 3.8. J48 decision tree predicting the GH subtypes highly associated with HMO degradation. A. LNFP I. **B.** LNFP II. **C.** LNFP III. The numbers in parentheses represent the total sample count and the instances that were classified incorrectly. The table displays information about the GH subtypes predicted to be involved in the enzymatic reaction of each substrate. Potential cleavage sites of the glycosidic linkage are indicated with an arrow.

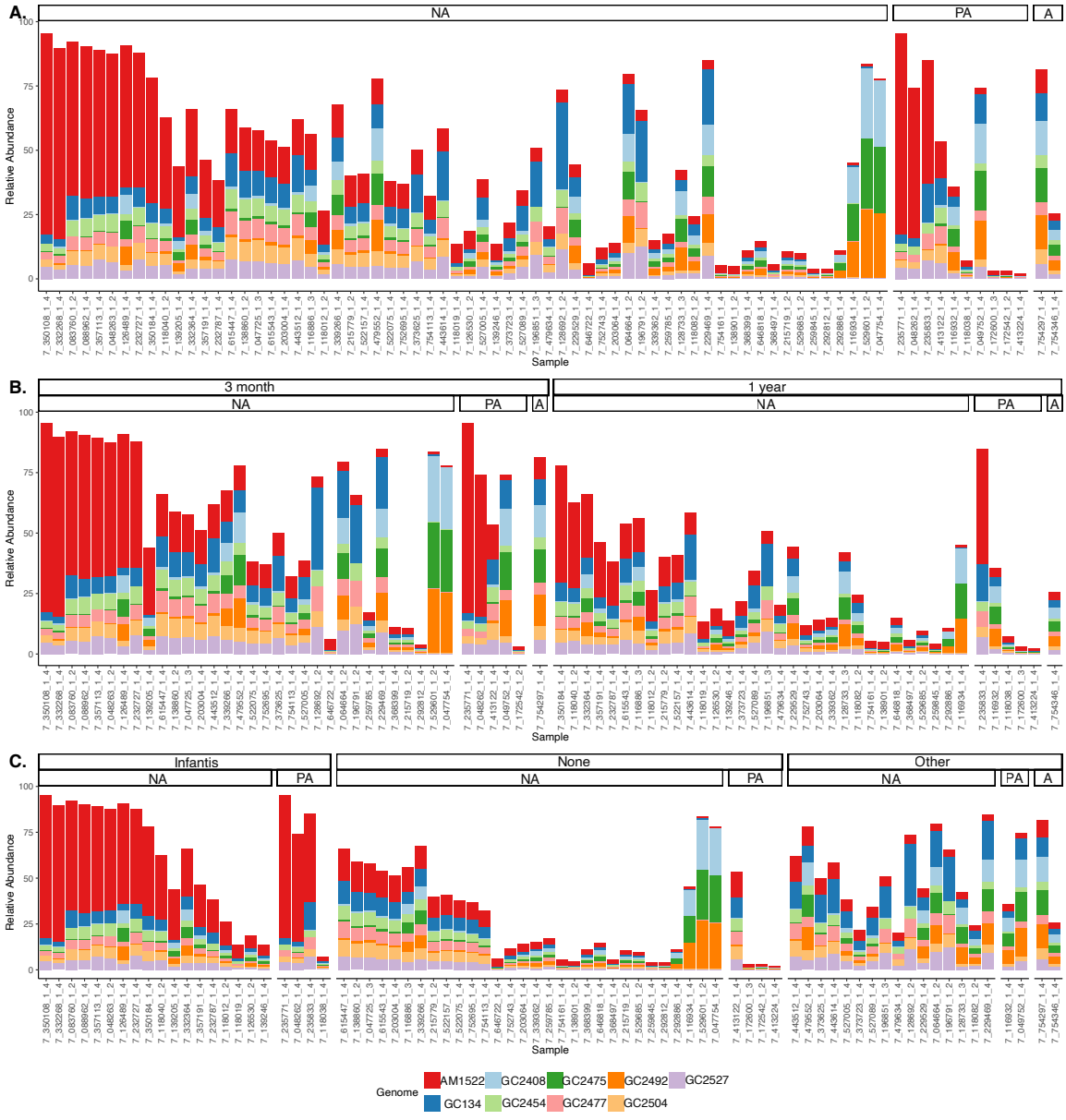


Figure 3.9. Relative abundance of nine *Bifidobacterium* strains in 76 metagenomic samples from the CHILD study. A. CHILD metagenomic samples are categorized based on asthma diagnosis at age 5. **B.** CHILD metagenomic samples are organized by based on the time of the visit where stool samples were collected. **C.** CHILD metagenomic samples are arranged according to StrainPhlAn output. “Infantis” indicates enrichment of *B. infantis*, “Other” represents the presence of other *B. longum* subspecies, and “None” indicates that no *B. longum* subspecies were detected by StrainPhlAn. NA; no asthma, A; asthma, PA; possible asthma.

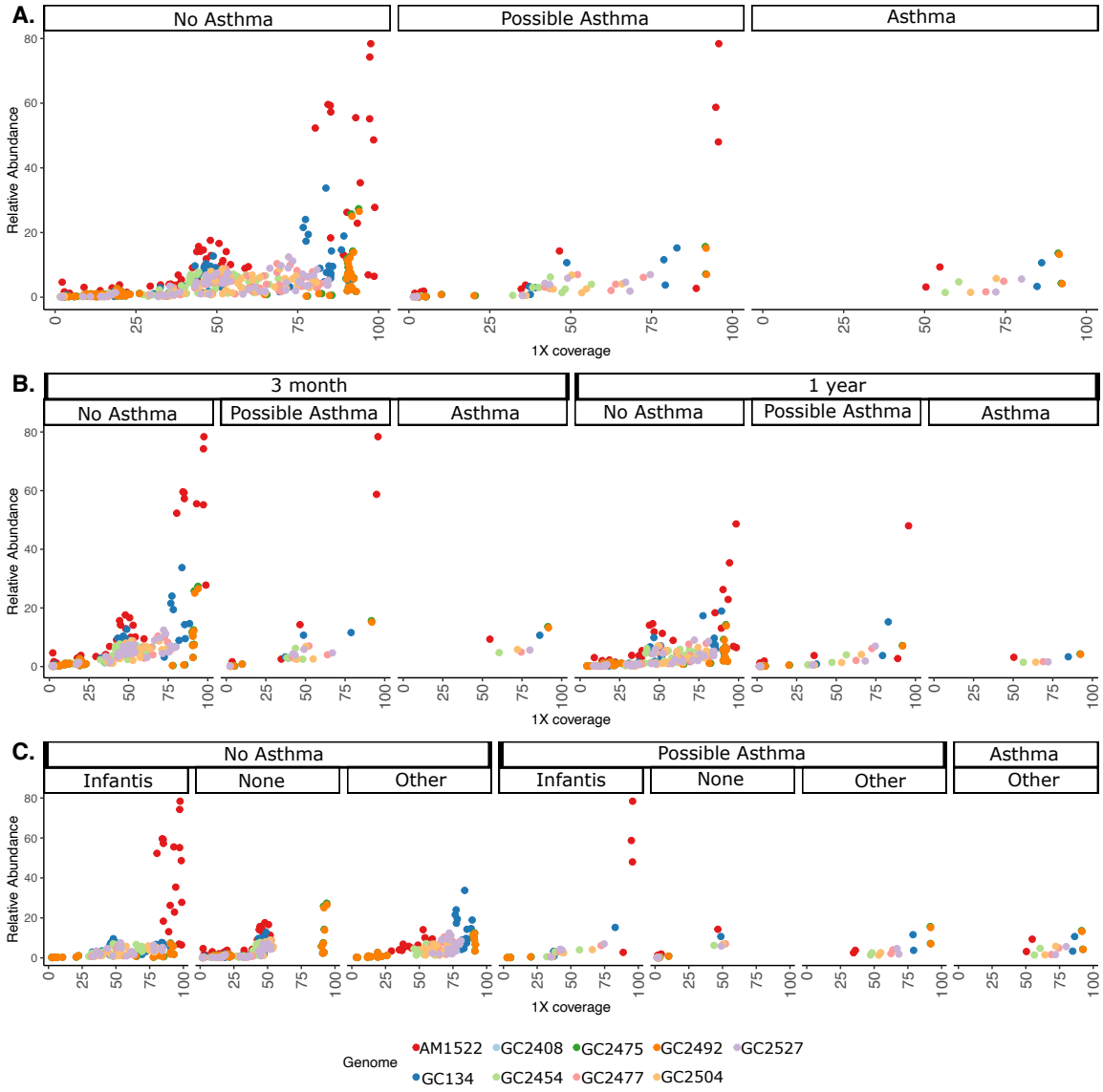


Figure 3.10. Relative abundance and genomic coverage of nine *Bifidobacterium* strains in 76 metagenomic samples from the CHILD study. Each dot represents the metagenomic sample. **A.** CHILD metagenomic samples are categorized based on asthma diagnosis at age 5. **B.** CHILD metagenomic samples are organized based on the time of the visit where stool samples were collected **C.** CHILD metagenomic samples are arranged according to StrainPhlAn output. "Infantis" indicates enrichment of *B. infantis*, "Other" represents the presence of other *B. longum* subspecies, and "None" indicates that no *B. longum* subspecies were detected by StrainPhlAn.

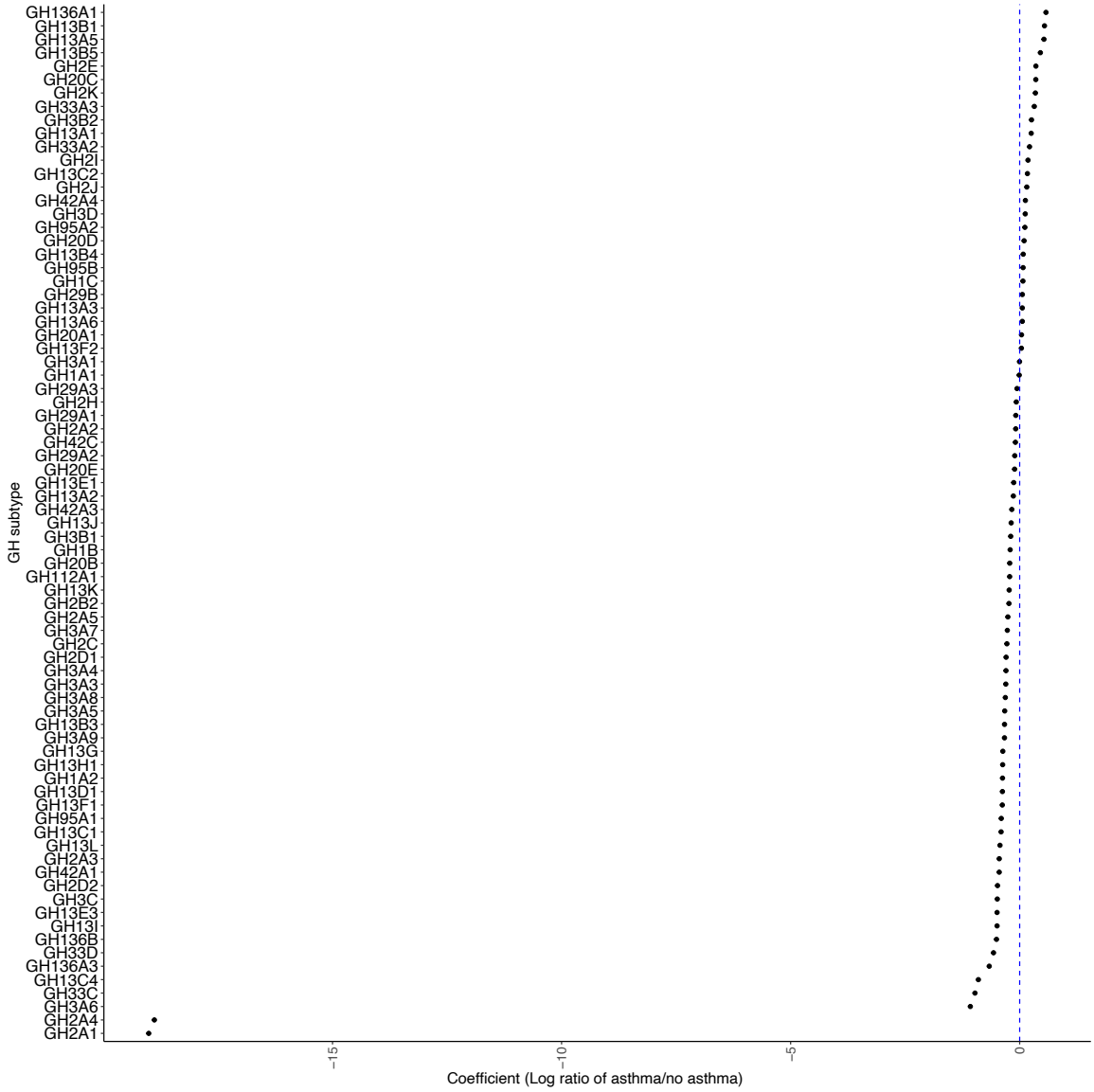


Figure 3.11. Changes in GH subtypes in individuals with asthma compared to those without asthma. X-axis represents the log fold change.

Table 3.1. Summary of non-bifidobacteria strains for HMO utilization assay.

Species	**Strain**
Bacteroides cellulosilyticus	GC234
Bacteroides ovatus	GC401
	GC137
Bacteroides xylanisolvens	GC232
Blautia luti	GC555
Blautia producta	GC553
Coprococcus eutactus	GC567
Enterococcus faecium	GC33
Fusicatenibacter saccharivorans	GC474
	GC313
Lactobacillus rhamnosus	GC38

Table 3.2. Nine selected strains for metagenomic read mapping.

Species	HMO degradation profile	Strain
Bifidobacterium longum subsp. infantis	Good	AM1522
Bifidobacterium longum	Good	GC2454
		GC2504
	Poor	GC134
		GC2477
		GC2527
Bifidobacterium bifidum	Good	GC2408
		GC2475
		GC2492

Table 3.3. CHILD Cohort study Metadata. Metagenomic data was provided by Dr. Charisse Peterson (UBC). The table presents metadata of 76 individuals, including information on study visits at 3 months and 1 year, as well as their asthma diagnosis at age 5.

Sample ID	Visit	StrainPhlAn	Asthma Diagnosis
7.754297.1.4	3 month	Other	Asthma
7.754346.1.4	1 year	Other	Asthma
7.083760.1.2	3 month	Infantis	No Asthma
7.088962.1.4	3 month	Infantis	No Asthma
7.357113.1.4	3 month	Infantis	No Asthma
7.350108.1.4	3 month	Infantis	No Asthma
7.232727.1.4	3 month	Infantis	No Asthma
7.048263.1.2	3 month	Infantis	No Asthma
7.332268.1.4	3 month	Infantis	No Asthma
7.128692.1.2	3 month	Other	No Asthma
7.196791.1.2	3 month	Other	No Asthma
7.350184.1.4	1 year	Infantis	No Asthma
7.229469.1.4	3 month	Other	No Asthma
7.118040.1.2	1 year	Infantis	No Asthma
7.196851.1.3	1 year	Other	No Asthma
7.064664.1.2	3 month	Other	No Asthma
7.126489.1.4	3 month	Infantis	No Asthma
7.443614.1.4	1 year	Other	No Asthma
7.332364.1.4	1 year	Infantis	No Asthma
7.357191.1.4	1 year	Infantis	No Asthma
7.232787.1.4	1 year	Infantis	No Asthma
7.373625.1.4	3 month	Other	No Asthma
7.527089.1.4	1 year	Other	No Asthma
7.139205.1.4	3 month	Infantis	No Asthma
7.479634.1.4	1 year	Other	No Asthma
7.229529.1.4	1 year	Other	No Asthma
7.373723.1.4	1 year	Other	No Asthma
7.527005.1.4	3 month	Other	No Asthma
7.118012.1.2	1 year	Infantis	No Asthma
7.118082.1.2	1 year	Other	No Asthma
7.128733.1.3	1 year	Other	No Asthma
7.443512.1.4	3 month	Other	No Asthma
7.118019.1.4	1 year	Infantis	No Asthma
7.126530.1.2	1 year	Infantis	No Asthma
7.139246.1.4	1 year	Infantis	No Asthma
7.479552.1.4	3 month	Other	No Asthma
7.339266.1.4	3 month	None	No Asthma
7.339362.1.4	1 year	None	No Asthma
7.368399.1.4	3 month	None	No Asthma
7.368497.1.4	1 year	None	No Asthma
7.138860.1.2	3 month	None	No Asthma
7.138901.1.2	1 year	None	No Asthma
7.215719.1.2	3 month	None	No Asthma
7.215779.1.2	1 year	None	No Asthma

7.522075.1.4	3 month	None	No Asthma
7.522157.1.4	1 year	None	No Asthma
7.615447.1.4	3 month	None	No Asthma
7.615543.1.4	3 month	None	No Asthma
7.752695.1.4	3 month	None	No Asthma
7.752743.1.4	1 year	None	No Asthma
7.754113.1.4	3 month	None	No Asthma
7.754161.1.4	1 year	None	No Asthma
7.529601.1.2	3 month	None	No Asthma
7.529685.1.2	1 year	None	No Asthma
7.047725.1.3	3 month	None	No Asthma
7.116886.1.3	1 year	None	No Asthma
7.047754.1.4	3 month	None	No Asthma
7.116934.1.4	1 year	None	No Asthma
7.203004.1.4	3 month	None	No Asthma
7.203064.1.4	1 year	None	No Asthma
7.259785.1.4	3 month	None	No Asthma
7.259845.1.4	1 year	None	No Asthma
7.292812.1.4	3 month	None	No Asthma
7.292886.1.4	1 year	None	No Asthma
7.646722.1.4	3 month	None	No Asthma
7.646818.1.4	1 year	None	No Asthma
7.235771.1.4	3 month	Infantis	Possible Asthma
7.235833.1.4	1 year	Infantis	Possible Asthma
7.048262.1.4	3 month	Infantis	Possible Asthma
7.049752.1.4	3 month	Other	Possible Asthma
7.116932.1.4	1 year	Other	Possible Asthma
7.118038.1.4	1 year	Infantis	Possible Asthma
7.413122.1.4	3 month	None	Possible Asthma
7.413224.1.4	1 year	None	Possible Asthma
7.172542.1.2	3 month	None	Possible Asthma
7.172600.1.3	1 year	None	Possible Asthma

CHAPTER 4. Conclusion

While breastfeeding is associated with the infant gut microbiome and the risk of asthma development, the utilization of HMOs by microbes and their relationship with the microbiome is not yet fully understood (Dai et al., 2023; Dai et al., 2022). To address this knowledge gap, CHILD cohort study was designed to investigate the causational roles of gut microbiome in pediatric asthma. Using data from the CHILD cohort study, the study aimed to determine the bacterial genes that are associated with decreased risk of developing or exacerbating asthma symptoms that act though HMO utilization, rather than focusing on specific *Bifidobacterium* species.

In this thesis, I have applied both computational methods and functional assays to profile 118 *Bifidobacterium* strains' GH genes and assess their ability to utilize HMOs. I performed comparative genomics and phenotypic screening to select strains to evaluate their capacities to degrade HMOs. I observed that different *Bifidobacterium* strains have varying abilities to degrade different carbohydrate sources (Chapter 2). Comparative genomics analysis suggested that one bifidobacterial strain can possess multiple genes within the same GH family. I hypothesized that those GH genes have distinct substrate specificity and act on different glycosidic linkages. To explore this, I constructed SSNs for GHs

involved in HMO degradation and classified GH genes within the same family into subtypes.

Selected *Bifidobacterium* strains and a few non-*Bifidobacterium* strains were subjected to HMO utilization assay and glycoprofiling which measured how much HMOs were consumed at different time points (Chapter 3). This analysis indicated strain-specific differences in HMO degradation, highlighting the importance of studying diverse strains. The HMO degradation data were used to train a machine learning model and construct decision trees for nineteen HMOs. These decision trees provided insights into which GH subtypes are associated with the degradation of each HMO. While some GH subtypes could be linked to specific degradation reactions, the resolution of our data was insufficient to assign all the GH subtypes. Mass-spectrometry data showing degradation products will allow me to refine these assignments.

I used the HMO utilization data and selected six good and three poor HMO degraders for further analysis. We sought to identify GH subtype genes that are protective against asthma development by mapping the metagenomic data from the CHILD study to these strains and GH subtype genes (Chapter 3). I observed that *B. infantis* was present and relatively abundant among individuals with asthma phenotypes. Additionally, a decrease in the relative abundance of *B. infantis* over time was observed, coinciding with an increased diversity in the infant gut microbiota. Our metagenomic read mapping suggests that asthma

development is influenced by a combination of factors rather than a single *Bifidobacterium* strain or GH subtype gene.

This thesis was not able to identify specific GH subtype genes associated with asthma risk. However, both *in vitro* and *in silico* approaches were used to study the HMO utilization in *Bifidobacterium* strains and to assign the GH proteins to HMO degradation. The study also suggests that asthma development is not limited to the presence of specific bacterial taxa. Expanding the analysis to all the GH families instead of GH families related to HMO degradation and using larger datasets of the CHILD study for metagenomic read mapping will offer more precise predictions.

The positive correlation between breastfeeding and bifidobacteria in infants is mediated through HMO utilization (Dai et al., 2023). However, the mechanisms behind this reduced risk of developing asthma are not completely understood. It may be an indirect effect when HMOs promote the expansion of bifidobacteria resulting in colonization resistance and reduced infection (Ackerman et al., 2017). The expansion of the bifidobacteria may slow the development of a more diverse microbiome, potentially altering the pathways of immune development (Depner et al., 2020). Alternatively, HMO utilization may activate the pathways in bifidobacteria that act directly on host pathways. In support of this, metagenomic analysis of CHILD samples has identified pathways independent of HMO utilization that are enriched in infants with a reduced asthma risk (C Peterson, personal communication). To explore the latter mechanism and understand the

impact of HMO utilization on other processes in the bacteria, RNA-seq could be performed to identify genes and pathways induced or repressed during *in vitro* HMO consumption. This could be complemented by metabolomic analysis.

There were a number of limitations to this study. I had access to only one *B. infantis* strain, so expanding the strain collection for this subspecies would be beneficial. Although many *B. infantis* genomes are available for analysis, combining comparative genomics and SSN analysis of GH proteins with experimental data on HMO utilization was essential for my work. Therefore, I restricted the bioinformatics analysis to the strains that I had access to. Another limitation was the data on HMO utilization. I only had information on whether a specific HMO was degraded. Higher resolution data showing the breakdown products would have allowed for a more refined mapping of GH subtypes to substrates. Nonetheless, this work lays the foundation for further studies on HMO degradation by bifidobacteria and could ultimately lead to rationally designed synbiotics of specific *Bifidobacterium* strains and HMOs to improve early life microbiome development for infants that cannot be breastfed.

Numerous studies are investigating the relationship between breastfeeding and asthma risk. One such study, which used network analysis, explored how milk microbes impact childhood asthma and allergic sensitization in milk-fed children (Yi Fang et al., 2024). The study found that greater diversity in human milk microbiota (HMM) was associated with a lower risk of childhood asthma, while higher levels of *Lawsonella* were linked to increased allergic sensitization

(Yi Fang et al., 2024). The study also highlighted the role of host genetics in influencing HMM and its potential effects on childhood asthma and atopy (Yi Fang et al., 2024). This highlights that multiple factors contribute to the development of asthma in childhood, and it is crucial to address these aspects in future research.

Another recent study used fecal metagenomics to examine the development of gut microbiota in newborns and assess microbial priority effects (Shao et al., 2024). This research identified different gut microbiota patterns, each led by specific bacteria and shaped by factors such as the mother's age and ethnic background (Shao et al., 2024). One pattern dominated by *Enterococcus faecalis* displayed unstable microbiota and high pathogen levels, while another pattern led by *Bifidobacterium* species, like *B. longum* and especially *B. breve*, exhibited a more stable microbiota and better pathogen resistance (Shao et al., 2024). This is likely due to their ability to utilize HMOs from breast milk (Sela et al., 2008). Interestingly, while *B. infantis* is known for its specialization in HMO utilization, it was absent or present in only small proportion of the UK and other Western cohorts, suggesting a potential lack of natural reservoirs for this species in these populations (Shao et al., 2024; Taft et al., 2022). Instead, other *Bifidobacterium* species or strains, like *B. breve* or *B. longum*, have taken over this functional niche. Given the limited success of probiotic *B. infantis* strains in establishing themselves in the gut microbiota of newborns, it is crucial to thoroughly examine the functional characteristics of naturally occurring and stable primary colonizers,

such as *B. breve*, in future studies (Shao et al., 2024). When studying the link between breastfeeding, reduced asthma risk, and the ability of bifidobacterial strains to utilize HMOs, it is crucial not to focus solely on the presence or absence of *B. infantis*, given its low prevalence in Western countries. Instead, the emphasis should be on the genes responsible for driving the beneficial outcomes.

CHAPTER 5. Bibliography

Ackerman, D. L., Craft, K. M., & Townsend, S. D. (2017). Infant food applications of complex carbohydrates: Structure, synthesis, and function. In *Carbohydrate Research* (Vol. 437). https://doi.org/10.1016/j.carres.2016.11.007

Ackerman, D. L., Doster, R. S., Weitkamp, J. H., Aronoff, D. M., Gaddy, J. A., & Townsend, S. D. (2017). Human Milk Oligosaccharides Exhibit Antimicrobial and Antibiofilm Properties against Group B Streptococcus. *ACS Infectious Diseases*, *3*(8). https://doi.org/10.1021/acsinfecdis.7b00064

Ahearn-Ford, S., Berrington, J. E., & Stewart, C. J. (2022). Development of the gut microbiome in early life. In *Experimental Physiology* (Vol. 107, Issue 5). https://doi.org/10.1113/EP089919

Ahmadizar, F., Vijverberg, S. J. H., Arets, H. G. M., de Boer, A., Garssen, J., Kraneveld, A. D., & Maitland-van der Zee, A. H. (2017). Breastfeeding is associated with a decreased risk of childhood asthma exacerbations later in life. *Pediatric Allergy and Immunology*, *28*(7). https://doi.org/10.1111/pai.12760

Akay, H. K., Bahar Tokman, H., Hatipoglu, N., Hatipoglu, H., Siraneci, R., Demirci, M., Borsa, B. A., Yuksel, P., Karakullukcu, A., Kangaba, A. A., Sirekbasan, S., Aka, S., Mamal Torun, M., & Kocazeybek, B. S. (2014). The relationship between bifidobacteria and allergic asthma and/or allergic dermatitis: A prospective study of 0–3 years-old children in Turkey. *Anaerobe*, *28*, 98–103. https://doi.org/10.1016/J.ANAEROBE.2014.05.006

Alessandri, G., van Sinderen, D., & Ventura, M. (2021). The genus Bifidobacterium: from genomics to functionality of an important component of the mammalian gut microbiota. *Computational and Structural Biotechnology Journal*, *19*. https://doi.org/10.1016/j.csbj.2021.03.006

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2). https://doi.org/10.1093/bioinformatics/btu638

Argüeso, P., Woodward, A. M., & AbuSamra, D. B. (2021). The Epithelial Cell Glycocalyx in Ocular Surface Infection. In *Frontiers in Immunology* (Vol. 12). https://doi.org/10.3389/fimmu.2021.729260

Armendáriz-Ruiz, M., Rodríguez-González, J. A., Camacho-Ruíz, R. M., & Mateos-Díaz, J. C. (2018). Carbohydrate esterases: An overview. In *Methods in Molecular Biology* (Vol. 1835). https://doi.org/10.1007/978-1-4939-8672-9_2

Arrieta, M. C., Stiemsma, L. T., Dimitriu, P. A., Thorson, L., Russell, S., Yurist-Doutsch, S., Kuzeljevic, B., Gold, M. J., Britton, H. M., Lefebvre, D. L., Subbarao, P., Mandhane, P., Becker, A., McNagny, K. M., Sears, M. R., Kollmann, T., Mohn, W. W., Turvey, S. E., & Finlay, B. B. (2015). Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Science Translational Medicine*, *7*(307). https://doi.org/10.1126/scitranslmed.aab2271

Asakuma, S., Hatakeyama, E., Urashima, T., Yoshida, E., Katayama, T., Yamamoto, K., Kumagai, H., Ashida, H., Hirose, J., & Kitaoka, M. (2011). Physiology of consumption of human milk oligosaccharides by infant gut-associated bifidobacteria. *Journal of Biological Chemistry*, *286*(40). https://doi.org/10.1074/jbc.M111.248138

Asakuma, S., Yokoyama, T., Kimura, K., Watanabe, Y., Nakamura, T., Fukuda, K., & Urashima, T. (2010). Effect of Human Milk Oligosaccharides on Messenger Ribonucleic Acid Expression of Toll-like Receptor 2 and 4, and of MD2 in the Intestinal Cell Line HT-29. *Journal of Applied Glycoscience*, *57*(3). https://doi.org/10.5458/jag.57.177

Asher, I., & Pearce, N. (2014). Global burden of asthma among children. *International Journal of Tuberculosis and Lung Disease*, *18*(11). https://doi.org/10.5588/ijtld.14.0170

Ashida, H., Miyake, A., Kiyohara, M., Wada, J., Yoshida, E., Kumagai, H., Katayama, T., & Yamamoto, K. (2009). Two distinct α-L-fucosidases from Bifidobacterium bifidum are essential for the utilization of fucosylated milk oligosaccharides and glycoconjugates. *Glycobiology*, *19*(9). https://doi.org/10.1093/glycob/cwp082

Autran, C. A., Kellman, B. P., Kim, J. H., Asztalos, E., Blood, A. B., Spence, E. C. H., Patel, A. L., Hou, J., Lewis, N. E., & Bode, L. (2018). Human milk

oligosaccharide composition predicts risk of necrotising enterocolitis in preterm infants. *Gut*, *67*(6). https://doi.org/10.1136/gutjnl-2016-312819

Azad, M. B., Vehling, L., Lu, Z., Dai, D., Subbarao, P., Becker, A. B., Mandhane, P. J., Turvey, S. E., Lefebvre, D. L., Sears, M. R., Anand, S. S., Befus, A. D., Brauer, M., Brook, J. R., Chen, E., Cyr, M., Daley, D., Dell, S., Denburg, J. A., … To, T. (2017). Breastfeeding, maternal asthma and wheezing in the first year of life: A longitudinal birth cohort study. *European Respiratory Journal*, *49*(5). https://doi.org/10.1183/13993003.02019-2016

Ballard, O., & Morrow, A. L. (2013). Human Milk Composition. Nutrients and Bioactive Factors. In *Pediatric Clinics of North America* (Vol. 60, Issue 1). https://doi.org/10.1016/j.pcl.2012.10.002

Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T. G., Fan, J., Garmiri, P., da Costa Gonzales, L. J., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., … Zhang, J. (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, *51*(D1), D523–D531. https://doi.org/10.1093/NAR/GKAC1052

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

Bidart, G. N., Rodríguez-Díaz, J., Monedero, V., & Yebra, M. J. (2014). A unique gene cluster for the utilization of the mucosal and human milk-associated glycans galacto-N-biose and lacto-N-biose in Lactobacillus casei. *Molecular Microbiology*, *93*(3). https://doi.org/10.1111/mmi.12678

Bode, L. (2012). Human milk oligosaccharides: Every baby needs a sugar mama. In *Glycobiology* (Vol. 22, Issue 9). https://doi.org/10.1093/glycob/cws074

Bode, L. (2015). The functional biology of human milk oligosaccharides. In *Early Human Development* (Vol. 91, Issue 11). https://doi.org/10.1016/j.earlhumdev.2015.09.001

Bode, L., Kuhn, L., Kim, H. Y., Hsiao, L., Nissan, C., Sinkala, M., Kankasa, C., Mwiya, M., Thea, D. M., & Aldrovandi, G. M. (2012). Human milk

oligosaccharide concentration and risk of postnatal transmission of HIV through breastfeeding. *American Journal of Clinical Nutrition*, *96*(4). https://doi.org/10.3945/ajcn.112.039503

Boix-Amorós, A., Collado, M. C., Van't Land, B., Calvert, A., Le Doare, K., Garssen, J., Hanna, H., Khaleva, E., Peroni, D. G., Geddes, D. T., Kozyrskyj, A. L., Warner, J. O., & Munblit, D. (2019). Reviewing the evidence on breast milk composition and immunological outcomes. *Nutrition Reviews*, *77*(8), 541–556. https://doi.org/10.1093/NUTRIT/NUZ019

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114. https://doi.org/10.1093/BIOINFORMATICS/BTU170

Boraston, A. B., Bolam, D. N., Gilbert, H. J., & Davies, G. J. (2004). Carbohydrate-binding modules: Fine-tuning polysaccharide recognition. In *Biochemical Journal* (Vol. 382, Issue 3). https://doi.org/10.1042/BJ20040892

Bosheva, M., Tokodi, I., Krasnow, A., Pedersen, H. K., Lukjancenko, O., Eklund, A. C., Grathwohl, D., Sprenger, N., Berger, B., & Cercamondi, C. I. (2022). Infant Formula With a Specific Blend of Five Human Milk Oligosaccharides Drives the Gut Microbiota Development and Improves Gut Maturation Markers: A Randomized Controlled Trial. *Frontiers in Nutrition*, *9*. https://doi.org/10.3389/fnut.2022.920362

Bottacini, F., Morrissey, R., Esteban-Torres, M., James, K., Van Breen, J., Dikareva, E., Egan, M., Lambert, J., Van Limpt, K., Knol, J., O'Connell Motherway, M., & Van Sinderen, D. (2018). Comparative genomics and genotype-phenotype associations in Bifidobacterium breve. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-28919-4

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R Journal*, *9*(2). https://doi.org/10.32614/rj-2017-066

Cantarel, B. I., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., & Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Research*, *37*(SUPPL. 1). https://doi.org/10.1093/nar/gkn663

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics*, *36*(6), 1925–1927. https://doi.org/10.1093/bioinformatics/btz848

Cheng, K., Zhou, Y., & Neelamegham, S. (2017). DrawGlycan-SNFG: A robust tool to render glycans and glycopeptides with fragmentation information. *Glycobiology*, *27*(3). https://doi.org/10.1093/glycob/cww115

Cheng, L., Kiewiet, M. B. G., Groeneveld, A., Nauta, A., & de Vos, P. (2019). Human milk oligosaccharides and its acid hydrolysate LNT2 show immunomodulatory effects via TLRs in a dose and structure-dependent way. *Journal of Functional Foods*, *59*. https://doi.org/10.1016/j.jff.2019.05.023

Chia, L. W., Mank, M., Blijenberg, B., Bongers, R. S., Van Limpt, K., Wopereis, H., Tims, S., Stahl, B., Belzer, C., & Knol, J. (2021). Cross-feeding between Bifidobacterium infantis and Anaerostipes caccae on lactose and human milk oligosaccharides. *Beneficial Microbes*, *12*(1). https://doi.org/10.3920/BM2020.0005

Chichlowski, M., Shah, N., Wampler, J. L., Wu, S. S., & Vanderhoof, J. A. (2020). Bifidobacterium longum subspecies infantis (B. infantis) in pediatric nutrition: Current state of knowledge. In *Nutrients* (Vol. 12, Issue 6). https://doi.org/10.3390/nu12061581

Craft, K. M., & Townsend, S. D. (2018). The Human Milk Glycome as a Defense Against Infectious Diseases: Rationale, Challenges, and Opportunities. *ACS Infectious Diseases*, *4*(2). https://doi.org/10.1021/acsinfecdis.7b00209

Dai, D. L. Y., Petersen, C., Hoskinson, C., Bel, K. L. Del, Becker, A. B., Moraes, T. J., Mandhane, P. J., Finlay, B. B., Simons, E., Kozyrskyj, A. L., Patrick, D. M., Subbarao, P., Bode, L., Azad, M. B., & Turvey, S. E. (2023). Breastfeeding enrichment of B. longum subsp. infantis mitigates the effect of antibiotics on the microbiota and childhood asthma risk. *Med*, *0*(0). https://doi.org/10.1016/J.MEDJ.2022.12.002

Dai, R., Miliku, K., Gaddipati, S., Choi, J., Ambalavanan, A., Tran, M. M., Reyna, M., Sbihi, H., Lou, W., Parvulescu, P., Lefebvre, D. L., Becker, A. B., Azad, M. B., Mandhane, P. J., Turvey, S. E., Duan, Q., Moraes, T. J., Sears, M. R., & Subbarao, P. (2022). Wheeze trajectories: Determinants and outcomes in the CHILD Cohort Study. *Journal of Allergy and Clinical Immunology*, *149*(6). https://doi.org/10.1016/j.jaci.2021.10.039

den Dekker, H. T., Sonnenschein-van der Voort, A. M. M., Jaddoe, V. W. V., Reiss, I. K., de Jongste, J. C., & Duijts, L. (2016). Breastfeeding and asthma outcomes at the age of 6 years: The Generation R Study. *Pediatric Allergy and Immunology*, *27*(5). https://doi.org/10.1111/pai.12576

Depner, M., Taft, D. H., Kirjavainen, P. V., Kalanetra, K. M., Karvonen, A. M., Peschel, S., Schmausser-Hechfellner, E., Roduit, C., Frei, R., Lauener, R., Divaret-Chauveau, A., Dalphin, J. C., Riedler, J., Roponen, M., Kabesch, M., Renz, H., Pekkanen, J., Farquharson, F. M., Louis, P., … Ege, M. J. (2020). Maturation of the gut microbiome during the first year of life contributes to the protective farm effect on childhood asthma. *Nature Medicine*, *26*(11). https://doi.org/10.1038/s41591-020-1095-x

Derakhshani, H., Bernier, S. P., Marko, V. A., & Surette, M. G. (2020). Completion of draft bacterial genomes by long-read sequencing of synthetic genomic pools. *BMC Genomics*, *21*(1). https://doi.org/10.1186/s12864-020-06910-6

Differding, M. K., Benjamin-Neelon, S. E., Hoyo, C., Østbye, T., & Mueller, N. T. (2020). Timing of complementary feeding is associated with gut microbiota diversity and composition and short chain fatty acid concentrations over the first year of life. *BMC Microbiology*, *20*(1). https://doi.org/10.1186/s12866-020-01723-9

Donovan, B. M., Abreo, A., Ding, T., Gebretsadik, T., Turi, K. N., Yu, C., Ding, J., Dupont, W. D., Stone, C. A., Hartert, T. V., & Wu, P. (2020). Dose, Timing, and Type of Infant Antibiotic Use and the Risk of Childhood Asthma. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *70*(8), 1658. https://doi.org/10.1093/CID/CIZ448

Duar, R. M., Casaburi, G., Mitchell, R. D., Scofield, L. N. C., Ortega Ramirez, C. A., Barile, D., Henrick, B. M., & Frese, S. A. (2020). Comparative Genome Analysis of Bifidobacterium longum subsp. infantis Strains Reveals Variation in Human Milk Oligosaccharide Utilization Genes among Commercial Probiotics. *Nutrients 2020, Vol. 12, Page 3247*, *12*(11), 3247. https://doi.org/10.3390/NU12113247

Fabiano, V., Indrio, F., Verduci, E., Calcaterra, V., Pop, T. L., Mari, A., Zuccotti, G. V., Cokugras, F. C., Pettoello-Mantovani, M., & Goulet, O. (2021). Term infant formulas influencing gut microbiota: An overview. In *Nutrients* (Vol. 13, Issue 12). https://doi.org/10.3390/nu13124200

Fang, Z., Pan, T., Li, L., Wang, H., Zhu, J., Zhang, H., Zhao, J., Chen, W., & Lu, W. (2022). Bifidobacterium longum mediated tryptophan metabolism to improve atopic dermatitis via the gut-skin axis. *Gut Microbes*, *14*(1). https://doi.org/10.1080/19490976.2022.2044723

Ferrante, G., & La Grutta, S. (2018). The burden of pediatric asthma. *Frontiers in Pediatrics*, *6*, 394966. https://doi.org/10.3389/FPED.2018.00186/BIBTEX

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23). https://doi.org/10.1093/bioinformatics/bts565

Gallier, S., Vocking, K., Post, J. A., Van De Heijning, B., Acton, D., Van Der Beek, E. M., & Van Baalen, T. (2015). A novel infant milk formula concept: Mimicking the human milk fat globule structure. *Colloids and Surfaces B: Biointerfaces*, *136*. https://doi.org/10.1016/j.colsurfb.2015.09.024

Garrido, D., Ruiz-Moyano, S., Kirmiz, N., Davis, J. C., Totten, S. M., Lemay, D. G., Ugalde, J. A., German, J. B., Lebrilla, C. B., & Mills, D. A. (2016). A novel gene cluster allows preferential utilization of fucosylated milk oligosaccharides in Bifidobacterium longum subsp. longum SC596. *Scientific Reports 2016 6:1*, *6*(1), 1–18. https://doi.org/10.1038/srep35045

Han, S. M., Derraik, J. G. B., Binia, A., Sprenger, N., Vickers, M. H., & Cutfield, W. S. (2021). Maternal and Infant Factors Influencing Human Milk Oligosaccharide Composition: Beyond Maternal Genetics. In *Journal of Nutrition* (Vol. 151, Issue 6). https://doi.org/10.1093/jn/nxab028

Harmsen, H. J. M., Wildeboer-Veloo, A. C. M., Raangs, G. C., Wagendorp, A. A., Klijn, N., Bindels, J. G., & Welling, G. W. (2000). Analysis of intestinal flora development in breast-fed and formula-fed infants by using molecular identification and detection methods. *Journal of Pediatric Gastroenterology and Nutrition*, *30*(1). https://doi.org/10.1097/00005176-200001000-00019

Hegar, B., Wibowo, Y., Basrowi, R. W., Ranuh, R. G., Sudarmo, S. M., Munasir, Z., Atthiyah, A. F., Widodo, A. D., Supriatmo, Kadim, M., Suryawan, A., Diana, N. R., Manoppo, C., & Vandenplas, Y. (2019). The Role of Two Human Milk Oligosaccharides, 2′-Fucosyllactose and Lacto-N-Neotetraose, in Infant Nutrition. *Pediatric Gastroenterology, Hepatology & Nutrition*, *22*(4), 330. https://doi.org/10.5223/PGHN.2019.22.4.330

Henrick, B. M., Hutton, A. A., Palumbo, M. C., Casaburi, G., Mitchell, R. D., Underwood, M. A., Smilowitz, J. T., & Frese, S. A. (2018). Elevated Fecal pH Indicates a Profound Change in the Breastfed Infant Gut Microbiome Due to Reduction of Bifidobacterium over the Past Century . *MSphere*, *3*(2). https://doi.org/10.1128/MSPHERE.00041-18/ASSET/23EA8212-75D9-4541-8EC1-8EFC38B8B6DF/ASSETS/GRAPHIC/SPH0021824890002.JPEG

Henrissat, B. (1991). A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochemical Journal*, *280*(2). https://doi.org/10.1042/bj2800309

Henrissat, B., & Davies, G. J. (2000). Glycoside hydrolases and glycosyltransferases. Families, modules, and implications for genomics. *Plant Physiology*, *124*(4). https://doi.org/10.1104/pp.124.4.1515

Hidalgo-Cantabrana, C., Delgado, S., Ruiz, L., Ruas-Madiedo, P., Sánchez, B., & Margolles, A. (2017). Bifidobacteria and Their Health-Promoting Effects. *Microbiology Spectrum*, *5*(3). https://doi.org/10.1128/microbiolspec.bad-0010-2016

Holscher, H. D., Bode, L., & Tappenden, K. A. (2017). Human Milk Oligosaccharides Influence Intestinal Epithelial Cell Maturation in Vitro. *Journal of Pediatric Gastroenterology and Nutrition*, *64*(2). https://doi.org/10.1097/MPG.0000000000001274

Hu, M., Miao, M., Li, K., Luan, Q., Sun, G., & Zhang, T. (2023). Human milk oligosaccharide lacto-N-tetraose: Physiological functions and synthesis methods. In *Carbohydrate Polymers* (Vol. 316). https://doi.org/10.1016/j.carbpol.2023.121067

Ioannou, A., Knol, J., & Belzer, C. (2021a). Microbial Glycoside Hydrolases in the First Year of Life: An Analysis Review on Their Presence and Importance in Infant Gut. *Frontiers in Microbiology*, *12*, 1345. https://doi.org/10.3389/FMICB.2021.631282/BIBTEX

Ioannou, A., Knol, J., & Belzer, C. (2021b). Microbial Glycoside Hydrolases in the First Year of Life: An Analysis Review on Their Presence and Importance in Infant Gut. In *Frontiers in Microbiology* (Vol. 12). https://doi.org/10.3389/fmicb.2021.631282

James, K., Motherway, M. O. C., Bottacini, F., & Van Sinderen, D. (2016). Bifidobacterium breve UCC2003 metabolises the human milk

oligosaccharides lacto-N-tetraose and lacto-N-neo-tetraose through overlapping, yet distinct pathways. *Scientific Reports*, *6*. https://doi.org/10.1038/srep38560

Janeček, Š., & Svensson, B. (2022). How many α-amylase GH families are there in the CAZy database? *Amylase*, *6*(1), 1–10. https://doi.org/10.1515/AMYLASE-2022-0001

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873). https://doi.org/10.1038/s41586-021-03819-2

Kassambara, A. (2020). ggpubr: "ggplot2" based publication ready plots. R package version 0.2. *Https://CRAN.R-Project.Org/Package=ggpubr*.

Kavanaugh, D., O'Callaghan, J., Kilcoyne, M., Kane, M., Joshi, L., & Hickey, R. M. (2015). The intestinal glycome and its modulation by diet and nutrition. *Nutrition Reviews*, *73*(6). https://doi.org/10.1093/nutrit/nuu019

Kennedy, K. M., de Goffau, M. C., Perez-Muñoz, M. E., Arrieta, M. C., Bäckhed, F., Bork, P., Braun, T., Bushman, F. D., Dore, J., de Vos, W. M., Earl, A. M., Eisen, J. A., Elovitz, M. A., Ganal-Vonarburg, S. C., Gänzle, M. G., Garrett, W. S., Hall, L. J., Hornef, M. W., Huttenhower, C., … Walter, J. (2023). Questioning the fetal microbiome illustrates pitfalls of low-biomass microbial studies. In *Nature* (Vol. 613, Issue 7945). https://doi.org/10.1038/s41586-022-05546-8

Kennedy, K. M., Gerlach, M. J., Adam, T., Heimesaat, M. M., Rossi, L., Surette, M. G., Sloboda, D. M., & Braun, T. (2021). Fetal meconium does not have a detectable microbiota before birth. *Nature Microbiology*, *6*(7). https://doi.org/10.1038/s41564-021-00904-0

Kim, J. H., An, H. J., Garrido, D., German, J. B., Lebrilla, C. B., & Mills, D. A. (2013). Proteomic Analysis of Bifidobacterium longum subsp. infantis Reveals the Metabolic Insight on Consumption of Prebiotics and Host Glycans. *PLOS ONE*, *8*(2), e57535. https://doi.org/10.1371/JOURNAL.PONE.0057535

Kitaoka, M. (2012). Bifidobacterial enzymes involved in the metabolism of human milk oligosaccharides. *Advances in Nutrition*, *3*(3). https://doi.org/10.3945/an.111.001420

Klopp, A., Vehling, L., Becker, A. B., Subbarao, P., Mandhane, P. J., Turvey, S. E., Lefebvre, D. L., Sears, M. R., Daley, D., Silverman, F., Hayglass, K., Kobor, M., Kollmann, T., Brook, J., Ramsey, C., Macri, J., Sandford, A., Pare, P., Tebbutt, S., … Hystad, P. (2017). Modes of Infant Feeding and the Risk of Childhood Asthma: A Prospective Birth Cohort Study. *The Journal of Pediatrics*, *190*, 192-199.e2. https://doi.org/10.1016/J.JPEDS.2017.07.012

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., Angenent, L. T., & Ley, R. E. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(SUPPL. 1), 4578–4585. https://doi.org/10.1073/PNAS.1000081107/-/DCSUPPLEMENTAL

Kong, C., Elderman, M., Cheng, L., de Haan, B. J., Nauta, A., & de Vos, P. (2019). Modulation of Intestinal Epithelial Glycocalyx Development by Human Milk Oligosaccharides and Non-Digestible Carbohydrates. *Molecular Nutrition and Food Research*, *63*(17). https://doi.org/10.1002/mnfr.201900303

Koromyslova, A., Tripathi, S., Morozov, V., Schroten, H., & Hansman, G. S. (2017). Human norovirus inhibition by a human milk oligosaccharide. *Virology*, *508*. https://doi.org/10.1016/j.virol.2017.04.032

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

Kunz, C., Rudloff, S., Baier, W., Klein, N., & Strobel, S. (2000). Oligosaccharides in human milk: structural, functional, and metabolic aspects. *Annual Review of Nutrition*, *20*, 699–722. https://doi.org/10.1146/ANNUREV.NUTR.20.1.699

Lawson, M. A. E., O'Neill, I. J., Kujawska, M., Gowrinadh Javvadi, S., Wijeyesekera, A., Flegg, Z., Chalklen, L., & Hall, L. J. (2020). Breast milk-derived human milk oligosaccharides promote Bifidobacterium interactions within a single ecosystem. *ISME Journal*, *14*(2). https://doi.org/10.1038/s41396-019-0553-2

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296. https://doi.org/10.1093/NAR/GKAB301

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14). https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16). https://doi.org/10.1093/bioinformatics/btp352

Lin, A. E., Autran, C. A., Szyszka, A., Escajadillo, T., Huang, M., Godula, K., Prudden, A. R., Boons, G. J., Lewis, A. L., Doran, K. S., Nizet, V., & Bode, L. (2017). Human milk oligosaccharides inhibit growth of group B Streptococcus. *Journal of Biological Chemistry*, *292*(27). https://doi.org/10.1074/jbc.M117.789974

Lin, C., Lin, Y., Zhang, H., Wang, G., Zhao, J., Zhang, H., & Chen, W. (2022). Intestinal 'Infant-Type' Bifidobacteria Mediate Immune System Development in the First 1000 Days of Life. In *Nutrients* (Vol. 14, Issue 7). https://doi.org/10.3390/nu14071498

Linhardt, R. J., Galliher, P. M., & Cooney, C. L. (1987). Polysaccharide lyases. In *Applied Biochemistry and Biotechnology* (Vol. 12, Issue 2). https://doi.org/10.1007/BF02798420

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research*, *42*(D1). https://doi.org/10.1093/nar/gkt1178

Lordan, C., Roche, A. K., Delsing, D., Nauta, A., Groeneveld, A., MacSharry, J., Cotter, P. D., & van Sinderen, D. (2024). Linking human milk oligosaccharide metabolism and early life gut microbiota: bifidobacteria and beyond. *Microbiology and Molecular Biology Reviews : MMBR*, *88*(1). https://doi.org/10.1128/MMBR.00094-23

Luna, E., Parkar, S. G., Kirmiz, N., Hartel, S., Hearn, E., Hossine, M., Kurdian, A., Mendoza, C., Orr, K., Padilla, L., Ramirez, K., Salcedo, P., Serrano, E., Choudhury, B., Paulchakrabarti, M., Parker, C. T., Huynh, S., Cooper, K., & Flores, G. E. (2022). Utilization Efficiency of Human Milk Oligosaccharides

by Human-Associated Akkermansia Is Strain Dependent. *Applied and Environmental Microbiology*, *88*(1). https://doi.org/10.1128/AEM.01487-21

Lv, H., Zhang, L., Han, Y., Wu, L., & Wang, B. (2022). The Development of Early Life Microbiota in Human Health and Disease. In *Engineering* (Vol. 12). https://doi.org/10.1016/j.eng.2020.12.014

Marcobal, A., Barboza, M., Froehlich, J. W., Block, D. E., German, J. B., Lebrilla, C. B., & Mills, D. A. (2010). Consumption of human milk oligosaccharides by gut-related microbes. *Journal of Agricultural and Food Chemistry*, *58*(9). https://doi.org/10.1021/jf9044205

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., & Sonnenburg, J. L. (2011). Bacteroides in the infant gut consume milk oligosaccharides via mucus-utilization pathways. *Cell Host and Microbe*, *10*(5). https://doi.org/10.1016/j.chom.2011.10.007

Marriage, B. J., Buck, R. H., Goehring, K. C., Oliver, J. S., & Williams, J. A. (2015). Infants Fed a Lower Calorie Formula with 2 ′ FL Show Growth and 2 ′ FL Uptake Like Breast-Fed Infants. *Journal of Pediatric Gastroenterology and Nutrition*, *61*(6). https://doi.org/10.1097/MPG.0000000000000889

Marsden, R. L., & Orengo, C. A. (2008). The classification of protein domains. *Methods in Molecular Biology*, *453*. https://doi.org/10.1007/978-1-60327-429-6_5

Milani, C., Duranti, S., Bottacini, F., Casey, E., Turroni, F., Mahony, J., Belzer, C., Palacio, S. D., Montes, S. A., Mancabelli, L., Lugli, G. A., Rodriguez, J. M., Bode, L., Vos, W. de, Gueimonde, M., Margolles, A., Sinderen, D. van, & Ventura, M. (2017). The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. *Microbiology and Molecular Biology Reviews : MMBR*, *81*(4). https://doi.org/10.1128/MMBR.00036-17

Milani, C., Lugli, G. A., Duranti, S., Turroni, F., Bottacini, F., Mangifesta, M., Sanchez, B., Viappiani, A., Mancabelli, L., Taminiau, B., Delcenserie, V., Barrangou, R., Margolles, A., Sinderen, D. van, & Ventura, M. (2014). Genomic encyclopedia of type strains of the genus Bifidobacterium. *Applied and Environmental Microbiology*, *80*(20). https://doi.org/10.1128/AEM.02308-14

Milani, C., Turroni, F., Duranti, S., Lugli, G. A., Mancabelli, L., Ferrario, C., Van Sinderen, D., & Ventura, M. (2016). Genomics of the genus Bifidobacterium reveals species-specific adaptation to the glycan-rich gut environment. *Applied and Environmental Microbiology*, *82*(4). https://doi.org/10.1128/AEM.03500-15

Miliku, K., & Azad, M. B. (2018). Breastfeeding and the developmental origins of asthma: Current evidence, possible mechanisms, and future research priorities. In *Nutrients* (Vol. 10, Issue 8). https://doi.org/10.3390/nu10080995

Morgan, W. J., Stern, D. A., Sherrill, D. L., Guerra, S., Holberg, C. J., Guilbert, T. W., Taussig, L. M., Wright, A. L., & Martinez, F. D. (2005). Outcome of asthma and wheezing in the first 6 years of life follow-up through adolescence. *American Journal of Respiratory and Critical Care Medicine*, *172*(10). https://doi.org/10.1164/rccm.200504-525OC

Morrow, A. L., Ruiz-Palacios, G. M., Altaye, M., Jiang, X., Lourdes Guerrero, M., Meinzen-Derr, J. K., Farkas, T., Chaturvedi, P., Pickering, L. K., & Newburg, D. S. (2004). Human milk oligosaccharides are associated with protection against diarrhea in breast-fed infants. *Journal of Pediatrics*, *145*(3). https://doi.org/10.1016/j.jpeds.2004.04.054

Newburg, D. S., & Grave, G. (2014). Recent advances in human milk glycobiology. *Pediatric Research*, *75*(5). https://doi.org/10.1038/pr.2014.24

Oberg, N., Zallot, R., & Gerlt, J. A. (2023). EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools. *Journal of Molecular Biology*, *435*(14). https://doi.org/10.1016/j.jmb.2023.168018

O'Callaghan, A., & van Sinderen, D. (2016). Bifidobacteria and their role as members of the human gut microbiota. In *Frontiers in Microbiology* (Vol. 7, Issue JUN). https://doi.org/10.3389/fmicb.2016.00925

Oddy, W. H., Holt, P. G., Sly, P. D., Read, A. W., Landau, L. I., Stanley, F. J., Kendall, G. E., & Burton, P. R. (1999). Association between breast feeding and asthma in 6 year old children: Findings of a prospective birth cohort study. *British Medical Journal*, *319*(7213). https://doi.org/10.1136/bmj.319.7213.815

Oftedal, O. T. (2012). The evolution of milk secretion and its ancient origins. *Animal*, *6*(3). https://doi.org/10.1017/S1751731111001935

Palframan, R. J., Gibson, G. R., & Rastall, R. A. (2003). Carbohydrate preferences of Bifidobacterium species isolated from the human gut. *Current Issues in Intestinal Microbiology*, *4*(2).

Pandey, K. R., Naik, S. R., & Vakil, B. V. (2015). Probiotics, prebiotics and synbiotics- a review. In *Journal of Food Science and Technology* (Vol. 52, Issue 12). https://doi.org/10.1007/s13197-015-1921-1

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3). https://doi.org/10.1093/bioinformatics/bty633

Penders, J., Thijs, C., Van Den Brandt, P. A., Kummeling, I., Snijders, B., Stelma, F., Adams, H., Van Ree, R., & Stobberingh, E. E. (2007). Gut microbiota composition and development of atopic manifestations in infancy: The KOALA birth cohort study. *Gut*, *56*(5). https://doi.org/10.1136/gut.2006.100164

Perez-Muñoz, M. E., Arrieta, M. C., Ramer-Tait, A. E., & Walter, J. (2017). A critical assessment of the “sterile womb” and “in utero colonization” hypotheses: Implications for research on the pioneer infant microbiome. In *Microbiome* (Vol. 5, Issue 1). https://doi.org/10.1186/s40168-017-0268-4

Pokusaeva, K., Fitzgerald, G. F., & Van Sinderen, D. (2011). Carbohydrate metabolism in Bifidobacteria. In *Genes and Nutrition* (Vol. 6, Issue 3). https://doi.org/10.1007/s12263-010-0206-6

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE*, *5*(3), e9490. https://doi.org/10.1371/JOURNAL.PONE.0009490

Puccio, G., Alliet, P., Cajozzo, C., Janssens, E., Corsello, G., Sprenger, N., Wernimont, S., Egli, D., Gosoniu, L., & Steenhout, P. (2017). Effects of infant formula with human milk oligosaccharides on growth and morbidity: A randomized multicenter trial. *Journal of Pediatric Gastroenterology and Nutrition*, *64*(4). https://doi.org/10.1097/MPG.0000000000001520

Qin, Z., Yang, S., Zhao, L., You, X., Yan, Q., & Jiang, Z. (2017). Catalytic mechanism of a novel glycoside hydrolase family 16 “elongating” β-transglycosylase. *Journal of Biological Chemistry*, *292*(5). https://doi.org/10.1074/jbc.M116.762419

Rabiu, B. A., Jay, A. J., Gibson, G. R., & Rastall, R. A. (2001). Synthesis and Fermentation Properties of Novel Galacto-Oligosaccharides by β-Galactosidases from Bifidobacterium Species. *Applied and Environmental Microbiology*, *67*(6). https://doi.org/10.1128/AEM.67.6.2526-2530.2001

Rozewicki, J., Li, S., Amada, K. M., Standley, D. M., & Katoh, K. (2019). MAFFT-DASH: Integrated protein sequence and structural alignment. *Nucleic Acids Research*, *47*(W1). https://doi.org/10.1093/nar/gkz342

Ruiz-Moyano, S., Totten, S. M., Garrido, D. A., Smilowitz, J. T., Bruce German, J., Lebrilla, C. B., & Mills, D. A. (2013). Variation in consumption of human milk oligosaccharides by infant gut-associated strains of bifidobacterium breve. *Applied and Environmental Microbiology*, *79*(19), 6040–6049. https://doi.org/10.1128/AEM.01843-13/SUPPL_FILE/ZAM999104746SO1.PDF

Ruiz-Palacios, G. M., Cervantes, L. E., Ramos, P., Chavez-Munguia, B., & Newburg, D. S. (2003). Campylobacter jejuni binds intestinal H(O) antigen (Fucα1, 2Galβ1, 4GlcNAc), and fucosyloligosaccharides of human milk inhibit its binding and infection. *Journal of Biological Chemistry*, *278*(16). https://doi.org/10.1074/jbc.M207744200

Saito, Y., Shigehisa, A., Watanabe, Y., Tsukuda, N., Moriyama-Ohara, K., Hara, T., Matsumoto, S., Tsuji, H., & Matsuki, T. (2020). Multiple Transporters and Glycoside Hydrolases Are Involved in Arabinoxylan-Derived Oligosaccharide Utilization in Bifidobacterium pseudocatenulatum. *Applied and Environmental Microbiology*, *86*(24). https://doi.org/10.1128/AEM.1782-20

Sarkar, A., Yoo, J. Y., Dutra, S. V. O., Morgan, K. H., & Groer, M. (2021). The association between early-life gut microbiota and long-term health and diseases. *Journal of Clinical Medicine*, *10*(3). https://doi.org/10.3390/jcm10030459

Schwengers, O., Jelonek, L., Dieckmann, M. A., Beyvers, S., Blom, J., & Goesmann, A. (2021). Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial Genomics*, *7*(11). https://doi.org/10.1099/MGEN.0.000685

Sela, D. A., Chapman, J., Adeuya, A., Kim, J. H., Chen, F., Whitehead, T. R., Lapidus, A., Rokhsar, D. S., Lebrilla, C. B., German, J. B., Price, N. P., Richardson, P. M., & Mills, D. A. (2008). The genome sequence of Bifidobacterium longum subsp. infantis reveals adaptations for milk utilization

within the infant microbiome. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(48), 18964–18969. https://doi.org/10.1073/PNAS.0809584105/SUPPL_FILE/0809584105SI.PDF

Shani, G., Hoeflinger, J. L., Heiss, B. E., Masarweh, C. F., Larke, J. A., Jensen, N. M., Wickramasinghe, S., Davis, J. C., Goonatilleke, E., El-Hawiet, A., Nguyen, L., Klassen, J. S., Slupsky, C. M., Lebrilla, C. B., & Mills, D. A. (2022). Fucosylated Human Milk Oligosaccharide Foraging within the Species Bifidobacterium pseudocatenulatum Is Driven by Glycosyl Hydrolase Content and Specificity. *Applied and Environmental Microbiology*, *88*(2). https://doi.org/10.1128/AEM.01707-21/SUPPL_FILE/AEM.01707-21-S0001.PDF

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11). https://doi.org/10.1101/gr.1239303

Shao, Y., Garcia-Mauriño, C., Clare, S., Dawson, N. J. R., Mu, A., Adoum, A., Harcourt, K., Liu, J., Browne, H. P., Stares, M. D., Rodger, A., Brocklehurst, P., Field, N., & Lawley, T. D. (2024). Primary succession of Bifidobacteria drives pathogen resistance in neonatal microbiota assembly. *Nature Microbiology*. https://doi.org/10.1038/s41564-024-01804-9

Shaw, A. G., Cornwell, E., Sim, K., Thrower, H., Scott, H., Brown, J. C. S., Dixon, R. A., & Kroll, J. S. (2020). Dynamics of toxigenic Clostridium perfringens colonisation in a cohort of prematurely born neonatal infants. *BMC Pediatrics*, *20*(1). https://doi.org/10.1186/s12887-020-1976-7

Spicer, S. K., Gaddy, J. A., & Townsend, S. D. (2022). Recent advances on human milk oligosaccharide antimicrobial activity. *Current Opinion in Chemical Biology*, *71*. https://doi.org/10.1016/J.CBPA.2022.102202

Stearns, J. C., Simioni, J., Gunn, E., McDonald, H., Holloway, A. C., Thabane, L., Mousseau, A., Schertzer, J. D., Ratcliffe, E. M., Rossi, L., Surette, M. G., Morrison, K. M., & Hutton, E. K. (2017). Intrapartum antibiotics for GBS prophylaxis alter colonization patterns in the early infant gut microbiome of low risk infants. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-16606-9

Taft, D. H., Lewis, Z. T., Nguyen, N., Ho, S., Masarweh, C., Dunne-Castagna, V., Tancredi, D. J., Huda, M. N., Stephensen, C. B., Hinde, K., von Mutius, E.,

Kirjavainen, P. V., Dalphin, J. C., Lauener, R., Riedler, J., Smilowitz, J. T., German, J. B., Morrow, A. L., & Mills, D. A. (2022). Bifidobacterium Species Colonization in Infancy: A Global Cross-Sectional Comparison by Population History of Breastfeeding. *Nutrients*, *14*(7). https://doi.org/10.3390/nu14071423

Tan, J., McKenzie, C., Potamitis, M., Thorburn, A. N., Mackay, C. R., & Macia, L. (2014). The Role of Short-Chain Fatty Acids in Health and Disease. In *Advances in Immunology* (Vol. 121). https://doi.org/10.1016/B978-0-12-800100-4.00003-9

Thomson, P., Medina, D. A., & Garrido, D. (2018a). Human milk oligosaccharides and infant gut bifidobacteria: Molecular strategies for their utilization. *Food Microbiology*, *75*, 37–46. https://doi.org/10.1016/J.FM.2017.09.001

Thomson, P., Medina, D. A., & Garrido, D. (2018b). Human milk oligosaccharides and infant gut bifidobacteria: Molecular strategies for their utilization. *Food Microbiology*, *75*, 37–46. https://doi.org/10.1016/J.FM.2017.09.001

Thursby, E., & Juge, N. (2017). Introduction to the human gut microbiota. In *Biochemical Journal* (Vol. 474, Issue 11). https://doi.org/10.1042/BCJ20160510

Tom Michael Mitchell. (1997). Machine Learning textbook. In *McGraw Hill*.

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., Frost, S. D. W., Corander, J., Bentley, S. D., & Parkhill, J. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology*, *21*(1), 1–21. https://doi.org/10.1186/S13059-020-02090-4/FIGURES/7

Trompette, A., Gollwitzer, E. S., Yadava, K., Sichelstiel, A. K., Sprenger, N., Ngom-Bru, C., Blanchard, C., Junt, T., Nicod, L. P., Harris, N. L., & Marsland, B. J. (2014). Gut microbiota metabolism of dietary fiber influences allergic airway disease and hematopoiesis. *Nature Medicine*, *20*(2). https://doi.org/10.1038/nm.3444

Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C., & Segata, N. (2017). Microbial strain-level population structure & genetic diversity from metagenomes. *Genome Research*, *27*(4). https://doi.org/10.1101/gr.216242.116

Turfkruyer, M., & Verhasselt, V. (2015). Breast milk and its impact on maturation of the neonatal immune system. In *Current Opinion in Infectious Diseases* (Vol. 28, Issue 3). https://doi.org/10.1097/QCO.0000000000000165

Turroni, F., Peano, C., Pass, D. A., Foroni, E., Severgnini, M., Claesson, M. J., Kerr, C., Hourihane, J., Murray, D., Fuligni, F., Gueimonde, M., Margolles, A., de Bellis, G., O'Toole, P. W., van Sinderen, D., Marchesi, J. R., & Ventura, M. (2012). Diversity of bifidobacteria within the infant gut microbiota. *PLoS ONE*, *7*(5). https://doi.org/10.1371/journal.pone.0036957

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

Turroni, F., van Sinderen, D., & Ventura, M. (2022). Bifidobacteria: insights into the biology of a key microbial group of early life gut microbiota. In *Microbiome Research Reports* (Vol. 1, Issue 1). https://doi.org/10.20517/mrr.2021.02

Underwood, M. A., German, J. B., Lebrilla, C. B., & Mills, D. A. (2014). Bifidobacterium longum subspecies infantis: champion colonizer of the infant gut. *Pediatric Research 2015 77:1*, *77*(1), 229–235. https://doi.org/10.1038/pr.2014.156

Van Den Broek, L. A. M., Hinz, S. W. A., Beldman, G., Vincken, J. P., & Voragen, A. G. J. (2008). Bifidobacterium carbohydrases-their role in breakdown and synthesis of (potential) prebiotics. In *Molecular Nutrition and Food Research* (Vol. 52, Issue 1). https://doi.org/10.1002/mnfr.200700121

Vestby, L. K., Grønseth, T., Simm, R., & Nesse, L. L. (2020). Bacterial biofilm and its role in the pathogenesis of disease. In *Antibiotics* (Vol. 9, Issue 2). https://doi.org/10.3390/antibiotics9020059

Victora, C. G., Bahl, R., Barros, A. J. D., França, G. V. A., Horton, S., Krasevec, J., Murch, S., Sankar, M. J., Walker, N., Rollins, N. C., Allen, K., Dharmage, S., Lodge, C., Peres, K. G., Bhandari, N., Chowdhury, R., Sinha, B., Taneja, S., Giugliani, E., … Richter, L. (2016). Breastfeeding in the 21st century: Epidemiology, mechanisms, and lifelong effect. In *The Lancet* (Vol. 387, Issue 10017). https://doi.org/10.1016/S0140-6736(15)01024-7

Waidyatillake, N. T., Allen, K. J., Lodge, C. J., Dharmage, S. C., Abramson, M. J., Simpson, J. A., & Lowe, A. J. (2013). The impact of breastfeeding on lung

development and function: A systematic review. In *Expert Review of Clinical Immunology* (Vol. 9, Issue 12). https://doi.org/10.1586/1744666X.2013.851005

Walker, A. (2010). Breast Milk as the Gold Standard for Protective Nutrients. *Journal of Pediatrics*, *156*(2 SUPPL.). https://doi.org/10.1016/j.jpeds.2009.11.021

Ward, R. E., Niñonuevo, M., Mills, D. A., Lebrilla, C. B., & German, J. B. (2006). In Vitro Fermentation of Breast Milk Oligosaccharides by Bifidobacterium infantis and Lactobacillus gasseri. *Applied and Environmental Microbiology*, *72*(6), 4497. https://doi.org/10.1128/AEM.02515-05

Wardman, J. F., Bains, R. K., Rahfeld, P., & Withers, S. G. (2022). Carbohydrate-active enzymes (CAZymes) in the gut microbiome. *Nature Reviews Microbiology 2022 20:9*, *20*(9), 542–556. https://doi.org/10.1038/s41579-022-00712-1

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology*, *13*(6), e1005595. https://doi.org/10.1371/JOURNAL.PCBI.1005595

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. In *Media* (Vol. 35, Issue July).

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43). https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2019). dplyr: A Grammar of Data Manipulation. R package version. In *Media*.

Wiederschain, G. Ya. (2009). Essentials of glycobiology. *Biochemistry (Moscow)*, *74*(9). https://doi.org/10.1134/s0006297909090156

Wong, C. B., Sugahara, H., Odamaki, T., & Xiao, J. Z. (2018). Different physiological properties of human-residential and non-human-residential bifidobacteria in human health. *Beneficial Microbes*, *9*(1). https://doi.org/10.3920/BM2017.0031

Xiao, M., Chuan Zhang, Hui Duan, Arjan Narbad, Jianxin Zhao, Wei Chen, Qixiao Zhai, Leilei Yu, & Fengwei Tian. (2024). *Cross-feeding of bifidobacteria promotes intestinal homeostasis: a lifelong perspective on the host health.*

Xue, M., Dehaas, E., Chaudhary, N., O'Byrne, P., Satia, I., & Kurmi, O. P. (2021). Breastfeeding and risk of childhood asthma: a systematic review and meta-analysis. *ERJ Open Research*, *7*(4). https://doi.org/10.1183/23120541.00504-2021

Yang, I., Corwin, E. J., Brennan, P. A., Jordan, S., Murphy, J. R., & Dunlop, A. (2016). The infant microbiome: Implications for infant health and neurocognitive development. *Nursing Research*, *65*(1). https://doi.org/10.1097/NNR.0000000000000133

Yi Fang, Z., Stickley, S. A., Ambalavanan, A., Zhang, Y., Zacharias, A. M., Fehr, K., Moossavi, S., Petersen, C., Miliku, K., Mandhane, P. J., Simons, E., Moraes, T. J., Sears, M. R., Surette, M. G., Subbarao, P., Turvey, S. E., Azad, M. B., & Duan, Q. (2024). *Networks of human milk microbiota are associated with host genomics, childhood asthma, and allergic sensitization*. https://doi.org/10.1016/j.chom.2024.08.014

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1). https://doi.org/10.1111/2041-210X.12628

Zabel, B. E., Gerdes, S., Evans, K. C., Nedveck, D., Singles, S. K., Volk, B., & Budinoff, C. (2020). Strain-specific strategies of 2′-fucosyllactose, 3-fucosyllactose, and difucosyllactose assimilation by Bifidobacterium longum subsp. infantis Bi-26 and ATCC 15697. *Scientific Reports 2020 10:1*, *10*(1), 1–18. https://doi.org/10.1038/s41598-020-72792-z

Zallot, R., Oberg, N., & Gerlt, J. A. (2019). The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry*, *58*(41). https://doi.org/10.1021/acs.biochem.9b00735

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Busk, P. K., Xu, Y., & Yin, Y. (2018). dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, *46*(W1), W95–W101. https://doi.org/10.1093/NAR/GKY418

Zhang, S., Li, T., Xie, J., Zhang, D., Pi, C., Zhou, L., & Yang, W. (2021). Gold standard for nutrition: a review of human milk oligosaccharide and its effects on infant gut microbiota. In *Microbial Cell Factories* (Vol. 20, Issue 1). https://doi.org/10.1186/s12934-021-01599-y