Bifidobacterium-host-diet interactions

Magdalena Kujawska

A thesis submitted for the degree of Doctor of Philosophy

University of East Anglia

Quadram Institute Bioscience

Norwich, United Kingdom

November 2020

Abstract

Bacteria belonging to the genus *Bifidobacterium* are key members of the gut microbiota. They are widely distributed in the animal kingdom, with over 80 recognised species and subspecies, and a host range spanning from insects to mammals. *Bifidobacterium* are among the earliest colonisers of the human gastrointestinal tract and have been associated with health-promoting benefits. However, investigations of infant-associated *Bifidobacterium* across early-life changing dietary periods are lacking. In addition, there is limited information on the diversity and the saccharolytic properties of this important microbiota member in diverse animal hosts. Thus, in this work I sought to comprehensively explore human- and animal-associated *Bifidobacterium* strains using both genomic and phenotypic approaches.

Whole genome sequencing (WGS) and bioinformatic analyses were employed to examine a unique collection of *Bifidobacterium longum* strains (n=75) isolated from nine either exclusively breast- or formula-fed infants across their first 18 months, encompassing pre-weaning, weaning and post-weaning dietary stages, as well as a novel collection of animal-associated *Bifidobacterium* isolates and publicly available sequences recovered from a diverse range of hosts (n=433). These genomes were analysed either in combination or as discrete subsets to determine their genomic diversity and predicted functional properties related to carbohydrate metabolism.

To complement bioinformatic analyses, a subset of infant-associated *B. longum* isolates were characterised phenotypically using experimental approaches to determine their carbohydrate metabolism capabilities, which linked to genomic analysis. Glycan uptake analysis and proteomics resulted in the determination of the mechanisms employed by selected *B. longum* strains to metabolise different carbohydrates.

Bacterial isolation resulted in the recovery of a substantial collection of animal-associated *Bifidobacterium* isolates (over 100) and the identification of potential novel species. The results of the bioinformatic analysis indicated a highly diverse “open” pan-genome and an overall very broad repertoire of carbohydrate

utilisation genes that could be associated with the host diet. This work represents the largest phylogenetic and comparative genomic analysis of animal-associated *Bifidobacterium* isolates to date.

Overall, this work enhances our current understanding of genomic and phenotypic properties of *Bifidobacterium* and lays the foundation for subsequent in-depth research aiming at further assessment of animal and human-associated *Bifidobacterium* diversity, and their functional potential for both therapeutic and industrial applications.

Declaration

I hereby declare that the material presented in this thesis has resulted from my own work during my PhD. All work done in collaboration has been appropriately acknowledged and accredited in the methods section as well as in each of the results chapters.

Magdalena Kujawska

November 2020
Norwich, UK

Acknowledgements

First of all, I wish to express my deep and sincere gratitude to my primary supervisor Professor Lindsay J Hall (QIB, TUM) for giving me the opportunity to do research and learn bioinformatics, for her dynamism, vision and enthusiastic mentorship, creative research ideas and continuous guidance. She has been unwavering in her support and optimism, and never failed to offer advice and encouragement, especially at times when I felt discouraged and unsure of my own work. I am also indebted to my secondary supervisor Professor Rob Kingsley (QIB) who offered constructive criticism, guidance and ideas, all of which were essential to the successful completion of my PhD.

I would also like to thank all my present and former colleagues in the Hall lab for friendly and supportive attitude, which made the lab a truly fantastic environment to work in. Many thanks to Cho Zin, Ian, Zoe, Mel, Cristina, Raymond, Lukas, Charlotte, Lisa, Sarah, Shannah, Holly, Matthew, Gowri, Dia, Iliana, Nancy, Anne and Peter for sharing their knowledge and professional and personal wisdom with me, and for all the fun experiences we enjoyed during the lab outings, team building and outreach activities. Special thanks go to Shab, who introduced me into the world of bioinformatics and taught me the importance of good coffee – for this I am eternally grateful.

I am indebted to my collaborators who have contributed to my PhD research, and whose input was essential for the completion of this work: Professor Lesley Hoyles (Nottingham Trent University); Dr Anne McCartney (University of Reading); Dr Phillip Pope, Dr Sabina Leanti La Rosa (Norwegian University of Life Sciences); Dr Sarah Knowles, (The Royal Veterinary College); Ms Aura Raulo (University of Oxford); Dr Laima Baltrūnaitė (Nature Research Centre, Lithuania); Ms Sara Goatcher (Banham Zoo and Africa Alive).

I would also like to acknowledge the funding bodies that supported my research: the BBSRC Doctoral Training Partnership (DTP) programme, and the Wellcome Trust New Investigator Award awarded to Professor Lindsay J Hall.

I would like to express my eternal appreciation towards my parents – without their love and support over the years it would not have been possible for me to achieve my educational goals; and to my wonderful and loving partner Bartosz, who unwaveringly supported my professional choices and stood by me throughout my PhD. Finally, I would also like to extend special thanks to my two fantastic friends, Ann-Marie and Dan, who have been on the same PhD journey and helped to keep me going.

Table of contents

List of figures

List of tables

List of abbreviations

2'-FL	2’-fucosyllactose
ANI	Average nucleotide identity
AX	Arabinoxylans
AXOS	Arabinoxylo-oligosaccharides
BHI	Brain Heart Infusion
BifCOG	*Bifidobacterium*-specific clusters of orthologous genes
BLAST	Basic Local Alignment Search Tool
BLASTN	BLAST search using Nucleotide query
BLASTP	BLAST search using Protein query
CAZy	Carbohydrate Active Enzymes
CBM	Carbohydrate-binding modules
CE	Carbohydrate esterase
COG	Clusters of orthologous groups
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DC-SIGN	Dendritic cell-specific intercellular adhesion molecule-3-grabbing non-integrin
DNA	Deoxyribonucleic acid
DP	Degree of polymerisation
DTH	Delayed-type hypersensitivity
EDTA	Ethylene Diamine Tetra-acetic Acid
EFSA	European Food Safety Authority
EPS	Exopolysaccharide
ESPGHAN	European Society for Paediatric Gastroenterology, Hepatology and Nutrition
FAO	Food and Agricultural Organisation of the United Nations
FDA	U.S. Food and Drug Administration
FOS	Fructo-oligosaccharides
GH	Glycosyl hydrolase
GOS	Galacto-oligosaccharides
GT	Glycosyl transferase
HGT	Horizontal Gene Transfer
HMO	Human milk oligosaccharide
HPAEC	High-performance anion-exchange chromatography
IMO	Isomalto-oligosaccharides
ISAPP	International Scientific Association for Prebiotics and Probiotics
ITF	Inulin-type fructans
ITS	Internally Transcribed Spacer
LNB	Lacto-*N*-biose
LNnT	Lacto-*N*-neotetraose
LNT	Lacto-*N*-tetraose
MLST	Multilocus Sequence Typing

MRS	de Man, Rogosa and Sharpe
NBI	Norwich Bioscience Institutes
NCBI	National Centre for Biotechnology Information
NDC	Non-digestible carbohydrates
ORF	Open reading frame
PCR	Polymerase Chain Reaction
PEP-PTS	Phosphoenolpyruvate-phosphotransferase
pGTF	Priming glycosyl transferase
POS	Pectin oligomers
QIB	Quadram Institute Bioscience
RCA	Reinforced Clostridial Medium
rRNA	Ribosomal Ribonucleic Acid
SBP	Solute binding proteins
SCFAs	Short-chain fatty acids
scFOS	Short-chain fructo-oligosaccharides
TUG	Truly unique genes
TUM	Technical University of Munich (Technische Universität München)
WGS	Whole genome sequencing
WHO	World Health Organisation
XOS	Xylo-oligosaccharides

List of publications that have arisen from this PhD

Microbes, Human Milk, and Prebiotics
Kujawska M, Collado MC, Hall LJ. In: Koren O and Rautava S (Eds). The Human Microbiome in Early Life: Implications to Health and Disease. San Diego: Elsevier Inc./Academic Press, 2021: 197-222.

The kleboxymycin biosynthetic gene cluster is encoded by several species belonging to the *Klebsiella oxytoca* complex.
Shibu P, McCuaig F, McCartney AL, Kujawska M, Hall LJ, Hoyles L. bioRxiv, 2020; doi: https://doi.org/10.1101/2020.07.24.215400

Microbiota supplementation with *Bifidobacterium* and *Lactobacillus* modifies the preterm infant gut microbiota and metabolome: an observational study.
Alcon-Giner C, Dalby MJ, Caim S, Ketskemety J, Shaw A, Sim K, Lawson M, Kiu R, Leclaire C, Chalklen L, Kujawska M, Mitra D, Fardus-Reid F, Belteki, G, McColl K, Swann JR, Kroll JS, Clarke P, Hall LJ. Cell Reports Medicine, 2020. doi: 10.1016/j.xcrm.2020.100077

Succession of *Bifidobacterium longum* strains in response to the changing early-life nutritional environment reveals specific adaptations to distinct dietary substrates.
Kujawska M, Leanti La Rosa S, Pope PB, Hoyles L, McCartney AL, Hall LJ. iScience, 2020. doi: 10.1016/j.isci.2020.101368

Rapid MinION profiling of preterm microbiota and antimicrobial resistant pathogens.
Alcon-Giner C, Leggett RM, Heavens D, Caim S, Brook TC, Kujawska M, Hoyles L, Clarke P, Clark MD/Hall LJ. Nature Microbiology, 2019. doi:10.1038/s41564-019-0626-z

Breast milk-derived human milk oligosaccharides promote *Bifidobacterium* interactions within a single ecosystem.
Lawson MAE, O'Neill IJ, Kujawska M, Wijeyesekera A, Flegg Z, Chalklen L, Hall LJ. ISME J. (2019). doi:10.1038/s41396-019-0553-2

Chapter 1

General introduction

This chapter is a literature review that constitutes the background of this thesis. The contents of this chapter are primarily based on the published book chapter "Microbes, human milk, and prebiotics" in the book "The Human Microbiome in Early Life", of which I am the first author.

Kujawska M, Collado MC, Hall LJ. Microbes, Human Milk, and Prebiotics In: Koren O and Rautava S (Eds). The Human Microbiome in Early Life: Implications to Health and Disease. San Diego: Elsevier Inc./Academic Press, 2021: 197-222

1.1 *Bifidobacterium*: history and general features

The genus *Bifidobacterium* belongs to phylum Actinobacteria, order Bifidobacteriales, family *Bifidobacteriaceae (1).* Bifidobacteria were first isolated in 1899 from faeces of breast-fed children by Henri Tissier, a French paediatrician, and named *Bacillus bifidus* due to their "bifid" Y-shape (Figure 1.1) (*1-3*). Because of their physiological and morphological similarity to lactobacilli, bifidobacteria were first classified under the genus *Lactobacillus* (*1*). In 1960s, de Vries and Stouthamer (*4*) reported the absence of catabolic enzymes aldolase and glucose-6-phosphate dehydrogenase in bifidobacteria, and showed the presence of a specific catabolic route involving fructose-6-phosphate phosphoketolase in these organisms. This observation led to re-classification of bifidobacteria as a separate genus in the 8th edition of Bergey's Manual of Determinative Bacteriology (*3*). Currently, the genus *Bifidobacterium* encompasses 87 taxa (July 2020) (*5*).

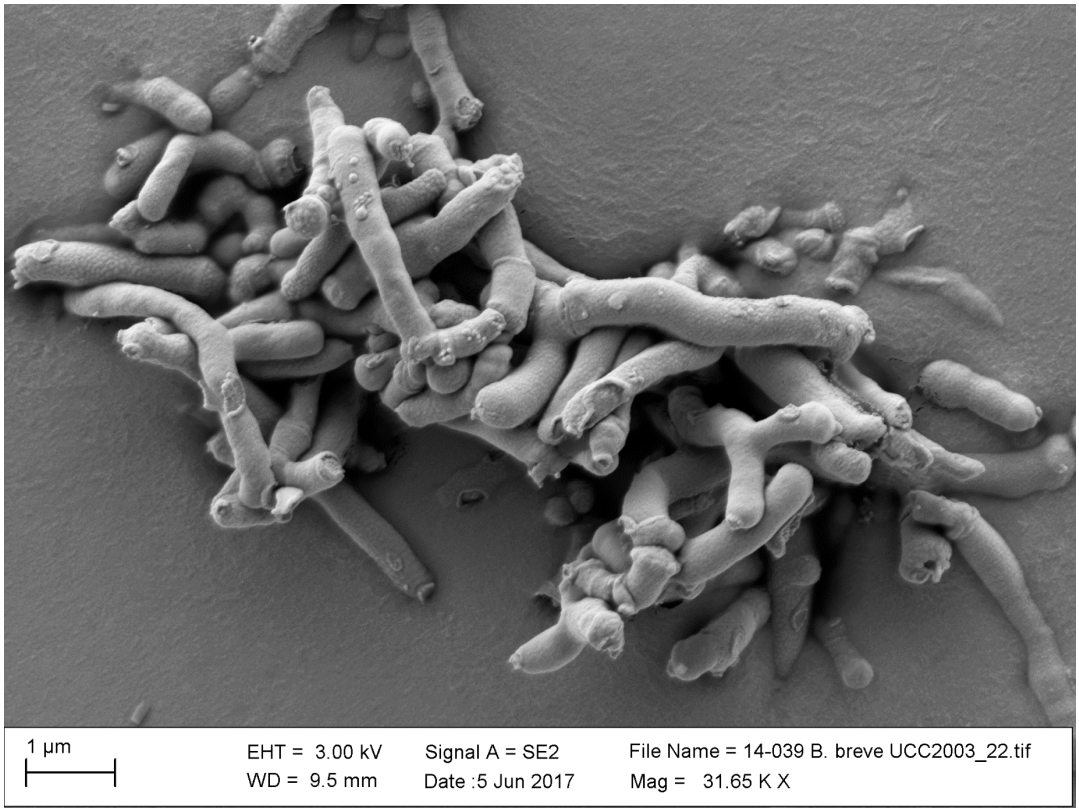


Figure 1.1 Electron microscopy image of vegetative cells of Bifidobacterium breve UCC2003. Courtesy of Mrs. Kathryn Cross (QIB).

Bifidobacteria are non-motile, non-spore-forming, non-gas-producing, and catalase-negative Gram-positive rods with high DNA G+C% content (42-67 mol%), suggested to be universally distributed among animals exhibiting parental care, including mammals, birds, reptiles, and social insects (*1, 6*). On solid media, bifidobacteria form smooth, convex colonies with entire edges, which are cream or white in colour and have soft consistency (*1*). The optimum growth temperature for the majority of *Bifidobacterium* species ranges between 37 and 41 ^{o}C, with the minimum and the maximum growth temperature reported at 25 and 45 ^{o}C, respectively (*1*). The literature suggests that members of the genus *Bifidobacterium* do not grow at pH below 4.5-5.0 and above 8.0, with the optimum pH for initial growth reported at 6.5 to 7.0 (*1*).

1.2 Isolation of bifidobacteria

Since the discovery of *Bifidobacterium*, various media formulations have been proposed and used for the isolation, culture and enumeration of these organisms from the gastrointestinal tract or faecal samples of humans and animals (*7*). It has been suggested that selection of an appropriate culture medium should be based on several specific parameters, namely: adequate supply of nutrients and growth substances, low oxidation–reduction potential, the final pH of the medium and the maintenance of the pH value during bacterial growth (*7*).

Hartemink and Rombouts (*8*) suggested that media suitable for bifidobacterial isolation and culture can be grouped into as many as five different classes, including non-selective media, media with elective carbohydrates, media with propionate, media with antibiotics and media with elective substance and/or low pH. Commercially available, non-selective media, such as reinforced clostridial agar (RCA) and de Man, Rogosa and Sharpe (MRS) have been shown to constitute excellent sources of nutrients required for *Bifidobacterium* growth and provide optimal growth conditions, including the adequate pH. The addition of L-cysteine aims at lowering the oxido-reduction potential in the medium and thus improves its

anaerobicity. Formulations with L-cysteine have been recommended for enumeration of *Bifidobacterium* from pure cultures (*9*).

Methods to isolate and enumerate bifidobacteria in samples with mixed bacterial populations have largely been based on the supplementation of known non-selective media with either a single carbon source or various selective agents that inhibit or reduce the growth of other bacteria. Raffinose is one example of a carbohydrate found to improve elective properties of both non-selective and selective media (*9*). The latter group contains formulations of varying complexity, whose selectivity is based either on antibiotics or on other appropriate ingredients, or both. For example, Beerens (*10*) established that the addition of propionate to Columbia agar at pH 5.0 stimulated the growth of bifidobacteria. Other studies found that Wilkins-Chalgren agar with added mupirocin was both more elective and selective for *Bifidobacterium* than that containing a combination of neomycin, paromomycin, nalidixic acid and lithium chloride (*11*). Overall, a large number of different types of media for isolation, cultivation and enumeration of *Bifidobacterium* have been developed over the years, however there is no standard formulation recommended for these purposes.

1.3 Approaches to bifidobacterial phylogeny

Until recently, the field of taxonomy and identification of new prokaryotic species relied on polyphasic characterisation based on a combination of phenotypic, chemotaxonomic and genotypic characteristics (*12*). The development of DNA amplification and sequencing techniques constituted an important step forward in determination of the taxonomic status of prokaryotes (*13*), and considerably increased the rate of the discovery of novel bacterial species (*14*). For many years, the analysis of the 16S ribosomal RNA gene was used as the primary tool for taxonomic assignment and phylogenetic trees, based on its presence in all bacteria, functional constancy and highly conserved structure (*15*). In 1994, Stackebrandt and Goebel (*16*) proposed a value of 97% 16S rRNA gene similarity as a threshold for the identification of new bacterial species. Since then, it has been demonstrated

that the discriminatory power of 16S rRNA gene sequences at this threshold could be insufficient at the species level (*17, 18*). More recently a cut-off value of 98.65% has been proposed (*19*), determined based on the pairwise comparison of 6787 prokaryotic genomes. However, independent studies have shown that for a number of bacterial species, including *Clostridium botulinum* and *Clostridium sporogenes*, *Rickettsia prowazekii* and *Rickettsia rickettsia*, or *Nocardia paucivorans* and *Nocardia brevicatena*, the inter-species 16S rRNA gene sequence similarity values are greater than 98.7 % (*20-22*). Thus, it has been postulated that the proposed 97% and 98.7% inter-species 16S rRNA gene sequence similarity thresholds should be used as indicators, rather than a definite tool for the classification of bacterial isolates (*23*).

Concerning the genus *Bifidobacterium*, the 16S rRNA gene allows for discrimination of most bifidobacterial species, however closely related taxa have been reported to exhibit high sequence homologies (*24, 25*). Thus, it is proposed that variation in the 16S rRNA gene alone, in the case of closely related bifidobacterial strains, is insufficient for clearly determining evolutionary distances, and the use of complementary phylogenetic markers for taxonomic differentiation has been advocated (*12, 24*). A phylogenetic approach using seven *Bifidobacterium* housekeeping genes (*clpC, dnaB, dnaG, dnaJ1, purF, rpoC, xfp*) has been found to allow a high level of discriminatory resolution between closely related bifidobacterial taxa, providing a robust means to infer phylogenetic relationships (*26*). In addition, the analysis of a non-coding 16S-23S internally transcribed spacer (ITS) has been shown to provide a greater level of resolution with regard to intraspecific phylogenetic relationships than the 16S rRNA gene (*24*). The procedure of concatenation has also been proposed as a measure of increasing efficacy and the robustness of phylogeny. The concatenated tree has been shown to allow for simultaneous phylogenetic inference at inter- and intraspecific level, and provides an increase in deep-node bootstrap values, thus increasing overall robustness (*24*).

Initial studies using a multilocus approach have identified six phylogenetic groups in the genus *Bifidobacterium*, namely the *Bifidobacterium adolescentis, Bifidobacterium asteroides, Bifidobacterium boum, B. longum, Bifidobacterium*

pseudolongum and *Bifidobacterium pullorum* groups (*26*). More recent investigations of 16S rRNA and 23S rRNA genes of type representatives of 47 described species and subspecies have recognised one more phylogenetic group, the *Bifidobacterium bifidum* group (*25*). The evaluation of DNA identity between analysed strains has shown that all bifidobacterial subspecies have the 16S rRNA and the 23S rRNA gene sequence homology above 97%. This analysis also revealed a high 16S rRNA and 23S rRNA gene sequence identity (>97%) between bifidobacterial taxa that are currently recognised as separate species (*25*).

1.4 Genomic characteristics of *Bifidobacterium*

The overall *Bifidobacterium* genome structure has recently been explored in studies that compared whole genome sequences of isolates representative of described bifidobacterial species, with the number of genomes included in these analyses ranging from 14 to 215 (*27-31*). The results revealed that bifidobacterial genomes ranged approximately from 1.63 Mb (*Bifidobacterium commune*) to 3.25 Mb (*Bifidobacterium biavatii*) in size, which corresponds to 1,237 and 2,557 predicted protein-encoding open reading frames (ORFs), respectively. Considering very close phylogenetic relationships between bifidobacterial taxa, it has been suggested, that the evolution of genomes has been driven by gene acquisition and/or loss events (*27, 32, 33*). Further analyses have suggested that those events may have contributed to the development of specific metabolic traits in bifidobacteria, allowing for transport and degradation of a vast range of carbohydrates (*27, 33*).

The functional annotation of ORFs has provided the basis for the *Bifidobacterium* pan-genome. The bacterial pan-genome represents both genes that appear to be conserved among bacteria in a particular taxonomic unit (core-genome) and genes that vary among them, either absent from the genome of at least one member of a taxonomic unit or unique to a single member of that taxonomic unit (*34, 35*). The size of pan-genome and recombination rates have been proposed to reflect differences in lifestyle and niche of different bacterial species (*34*). Considering the dynamic nature of bacterial populations and their tendency to evolve and exchange

genetic material, one limitation of pan-genome analysis is the difficulty of assessing whether a pan-genome is “closed” or not. In general, an “open” pan-genome of a specific bacterial group indicates that genomes in this group are evolving and diversify, whereas a “closed” pan-genome implies a low gene exchange (*36*).

The work of Milani et al. (*27*) laid the foundation for *Bifidobacterium* genomic analyses. The functional analysis of 47 genome sequences identified 18,181 *Bifidobacterium*-specific clusters of orthologous genes (BifCOGs), which represent the pan-genome, and 551 BifCOGs constituting the core-genome (Figure 1.2) (*27*). Overall, the majority of the conserved core genes have been determined to encode housekeeping functions and functions involved in the adaptation to or in the interaction with a particular environment. The authors of this study were primarily interested in carbohydrate utilisation by bifidobacteria, and in this context, it has been estimated that only 5.5% of the core-genome has functions associated with carbohydrate metabolism, whereas the carbohydrate metabolism functional family is the most represented in the pan-genome (13.7%). These results indicated the existence of selective pressure with regard to the acquisition and retention of genes for carbohydrate utilization by bifidobacteria (*27*).

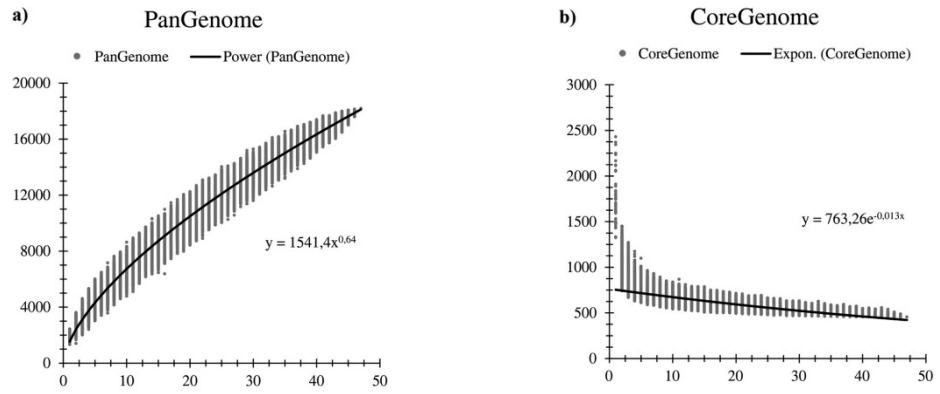


Figure 1.2 The pan-genome of genus Bifidobacterium.
The pan-genome (a) and core-genome (b) are represented as sizes of their gene pools versus the analysed 47 bifidobacterial genomes. The x axes represent the number of genomes, whereas the y axes represent the number of genes. Reproduced from Milani et al., 2014, Applied and Environmental Microbiology, DOI: 10.1128/AEM.02308-14 (27), with the authorisation from the American Society for Microbiology to republish the requested material in the doctoral thesis.

The pan-genome and the core-genome power trends can be visualised by plotting the pan-genome and core-genome sizes as functions of the number of analysed genomes. The results of such analysis have shown that the pan-genome trend line for bifidobacteria has not reached a plateau (Figure 1.2). Although the number of new genes discovered by sequential addition of new genome sequences has recently decreased, the addition of new genomes is expected to further reduce the number of new gene discoveries. This indicates the existence of an "open" pan-genome within the *Bifidobacterium* genus (*27*). Thus, it has been suggested that additional studies are needed in order to identify all genes present in the members of genus *Bifidobacterium*. In contrast, the trend line for the core gene set in bifidobacteria has been shown to reach a plateau, which suggests that the number of core genes is not expected to be significantly reduced by the addition of new genome sequences (*27*).

The pan-genome analysis has also determined the bifidobacterial variome, encompassing truly unique genes (TUGs) present in just one genome. The mean number of TUGs in bifidobacteria has been identified as 249, but a large deviation from the mean has been observed within the genus, with *Bifidobacterium indicum* and *Bifidobacterium cuniculi* having 47 and 595 TUGs, respectively. These observations have suggested a high degree of genome diversity between bifidobacterial species, which has been proposed to reflect individual adaptations to different environments. Over 14% of TUGs have been found to be involved in carbohydrate transport or degradation, thus it has been suggested that the analysis of unique genes may prove a useful tool in studying bifidobacterial metabolism of host- or diet-derived components (*27*).

The availability of complete genome sequences and pan-genomic analyses have facilitated a more robust phylogeny reconstruction. Conserved genes in bacterial core-genomes are relatively unlikely to experience horizontal gene transfer events, which makes them appropriate targets for phylogenetic inference (*37*). Concatenated protein sequences based on core-genomes can be used to create phylogenetic supertrees. Such analysis of 47 bifidobacterial genomes has revealed the discriminatory power of the concatenated protein tree to be significantly higher

than that observed for the tree based on 16S rRNA gene, with much higher bootstrap support (*25*). The increase in sequence length, as well as the use of protein-based sequences has resulted in a considerable increase in tree robustness. Thus, it has been advocated that the analysis based on bifidobacterial core-genome provides a more reliable phylogenetic method than the analysis of the 16S rRNA gene sequences (*25*).

To date, a limited number of studies attempting global genomic and phylogenetic analyses of members of genus *Bifidobacterium* have been published (*25, 27, 28, 30, 31, 33*). One caveat of these studies is that they either focused on a subset of genomes, or only included genomes of bifidobacterial type strains isolated from a limited host range. Some of the more recently recognised *Bifidobacterium* species were not represented in these analyses. The incorporation of new genome sequences obtained from additional host sources into these types of analysis could provide a more comprehensive and robust insight into the progression and functional characterisation of pan- and core-genomes of the genus. With carbohydrate metabolism being a key feature of bifidobacteria, the focus of a number of published genus-wide genomic comparisons has primarily been on functional gene groups involved in carbohydrate utilisation (*27, 33, 38*). However, more analyses are needed to encompass the diversity of the members of *Bifidobacterium* and assess the extent of their potential in terms of carbohydrate metabolism.

1.5 An overview of carbohydrate metabolism in bifidobacteria

Most simple sugars are absorbed or metabolised in the upper parts of the intestinal tract, whereas complex carbohydrates, for which the host lacks digestive capacity, are utilised by intestinal bacteria in the lower parts of the gut (*39*). The ability of commensal bacteria to degrade complex carbohydrates, such as dietary compounds (e.g. resistant starch, hemicellulose, glycogen), host-derived compounds (e.g. mucin, glycosphingolipids, chondroitin sulphate, hyaluronic acid), or carbon sources produced by other members of the gut microbial community has

been well established (*40-42*). It has been suggested that the amount and nature of complex carbohydrates in human diet directly impacts the metabolic activity and composition of the gastro-intestinal microbiota (*43*).

Bifidobacteria are saccharolytic and utilise carbohydrates as their sole source of carbon and energy. As such, they are believed to play a key role in carbohydrate metabolism in the colon, where they are most prevalent. Bifidobacterial genomes have been suggested to reflect adaptations of members of this group to the environment of the host gastro-intestinal tract (*33, 44-46*), and are characterised by the presence of genes that encode a variety of carbohydrate-modifying enzymes, e.g. glycosyl hydrolases, ABC transporters, and the components of PEP-PTS (phosphoenolpyruvate-phosphotransferase) system involved in the concomitant transport and phosphorylation of carbohydrates (*47*).

Phenotypic studies have confirmed that bifidobacteria can metabolise a wide range of complex carbohydrates, including host-derived gastric mucin, and plant-derived oligosaccharides, such as xylo-oligosaccharides, fructo-oligosaccharides, or pectin, but the bifidobacterial metabolic capacity for specific carbohydrates has been suggested to be species- and strain-dependent (*42*). Many of the characterised bifidobacteria can metabolise ribose, galactose, fructose, glucose, sucrose, maltose, melibiose and raffinose, but generally cannot utilise L-arabinose, rhamnose, *N*-acetylglucosamine, sorbitol, melezitose, trehalose, glycerol, xylitol and inulin (*42*). In general, gastro-intestinal bacteria degrade complex polymeric carbohydrates to low molecular weight oligosaccharides, which can subsequently be degraded to monosaccharides (*4, 42*). In bifidobacteria, these monosaccharides are converted to intermediates of a particular hexose fermentation pathway, termed the ‘‘bifid shunt’’, and ultimately converted to metabolic end products, which include lactate, short-chain fatty acids (SCFAs), e.g. acetate and formate, and other organic compounds (Figure 1.3) (*4, 42, 48*).

The key enzyme in the “bifid shunt” is fructose-6-phosphate phosphoketolase (EC 4.1.2.2). It has been shown to be present in all members of the family *Bifidobacteriaceae*, and is thus considered to be a taxonomic marker for members

of this group (*49*). Fermentation through the "bifid shunt" provides an advantage for bifidobacteria over, for example, lactic acid bacteria, as it yields 2.5 moles of ATP, 1.5 mole of acetate and 1 mole of lactate from every 1 mole of fermented glucose (*50*). The ratio of lactate to acetate may depend on the specific carbohydrate being metabolised, and on the bifidobacterial species. In addition, it has been shown that rapid fermentation of carbohydrates results in the production of substantial amounts of lactate, whereas acetate, formate and ethanol production increases when carbohydrates are degraded at a slower rate (*51*).

The produced SCFAs are believed to be beneficial to the host. They provide around 10% of the daily caloric requirement and are used as energy source by colonocytes and hepatocytes (*52, 53*). In addition, SCFAs modulate the development and function of intestinal epithelial cells and cells of the innate and adaptive immune system, including neutrophils, macrophages and T cells, through activation of G protein coupled receptors (*54*). Additionally, SCFAs stimulate adsorption of sodium and water in the colon and are known to induce enzymes promoting mucosal restitution (*53*). Furthermore, they can be used as co-substrates in the production of butyrate by other colonic bacteria through cross-feeding interactions (*55*). These kinds of interactions have been suggested to favour the co-existence of specific *Bifidobacterium* species and strains with other bifidobacteria and with butyrate-producing bacteria, e.g. *Faecalibacterium prausnitzii* and *Roseburia* species, in the colon (*38, 55*).

Approximately half of the genes functionally annotated as involved in carbohydrate metabolism in each bifidobacterial genome have been estimated to take part in carbohydrate uptake, as either ABC transporters, permeases or proton symporters (*56*). The uptake of most of the complex sugars is thought to be facilitated through the ABC transporter system, and only a small number of the carbohydrates fermented by bifidobacteria are believed to be transported via PEP-PTS system (*57*). Following internalization, complex carbohydrates can be hydrolysed, phosphorylated, deacetylated and/or transglycosylated by specific intracellular enzymes (*56*).

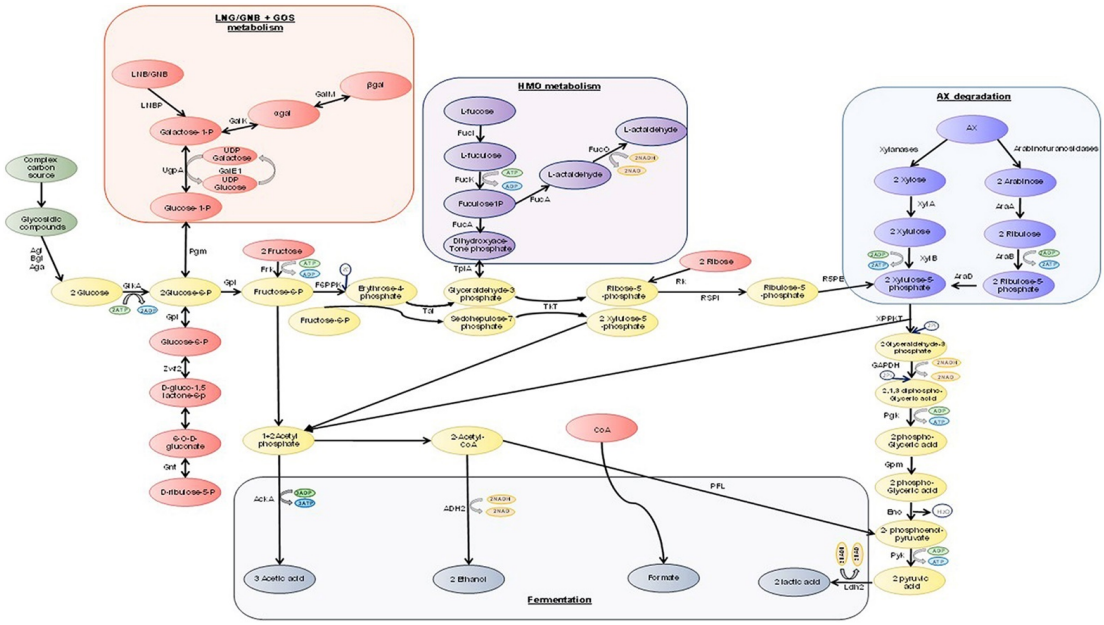


Figure 1.3 A representation of carbohydrate degradation through the "bifid shunt" in bifidobacteria. Reproduced from O'Callaghan and van Sinderen, 2016, Frontiers in Microbiology, DOI: 10.3389/fmicb.2016.00925 (45), with the permission from Frontiers under the Creative Commons Attribution License.

Glycosyl hydrolases (GHs) appear to be the most prevalent and critical group of carbohydrate-modifying enzymes for bifidobacteria (*38*). In general, their mode of action involves hydrolysing the glycosidic bond between two or more sugars, or between a sugar and a non-sugar moiety in the presence of water (*42, 48*). However, when a high concentration of sugar is used in the reaction, specific GHs, termed retaining glycosyl hydrolases, can use the carbohydrate molecule as an acceptor molecule instead of water, which results in the exchange of the sugar residues, and can lead to formation of new oligosaccharides with a higher degree of polymerization (transglycosylation reaction) (Figure 1.4) (*42*).

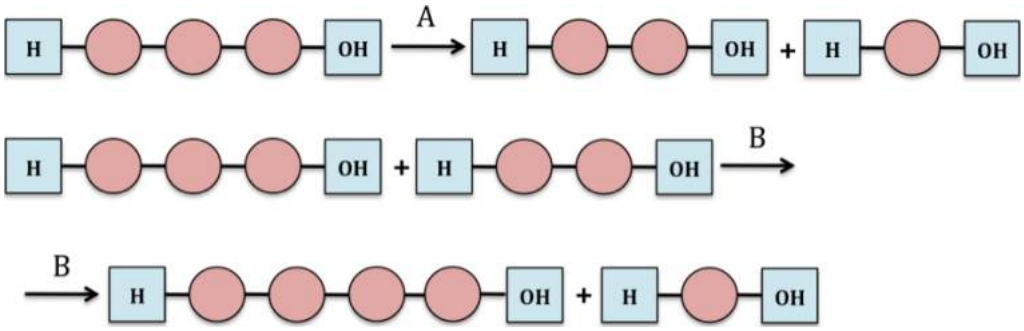


Figure 1.4 A representation of hydrolysis (A) and transglycosylation (B) reactions performed by glycosyl hydrolases.
Reproduced from Pokusaeva et al., 2011, Genes & Nutrition, DOI: 10.1007/s12263-010-0206-6 (39), with the permission obtained from Springer Nature through Copyright Clearance Center's RightsLink® service.

The analysis of the bifidobacterial pan-genome, based on 47 *Bifidobacterium* type strains, and subsequent classification according to the Carbohydrate Active Enzymes (CAZy) system have revealed that members of the genus *Bifidobacterium* encompass one of the largest predicted glycobiomes among known gastro-intestinal commensals (*38*). In total, 3,385 genes have been identified that represent carbohydrate-active enzymes, including 57 families of glycosyl hydrolases (GHs), 13 families of glycosyl transferases (GTs), and 7 families of carbohydrate esterases (CEs). The genus *Bifidobacterium* have been shown to possess an extensive repertoire of enzymes belonging to GH13, GH43, GH3 and GH51 families (2.0 fold-, 2.6 fold-, 5.8 fold-, 7.0 fold more, respectively, compared to the average GH arsenal of the gut microbiome), along with *Bacteroides* spp. (*Bacteroidales* family) and members of the *Clostridiales* family (*38*) . The enzymes belonging to GH13 family of glycosyl hydrolases have been found to dominate in the bifidobacterial glycobiome. These enzymes have the capacity to hydrolyse a wide range of complex plant-derived carbohydrates, such as starch and related substrates (e.g. amylose and maltodextrin), as well as palatinose, stachyose, raffinose, and melibiose, which are commonly present in the adult mammalian diet (*58*).

In addition, the bifidobacterial glycobiome has been shown to comprise members of GH families crucial for host glycan breakdown, for example those of families GH33 and GH34 (exo-sialidases), family GH29, which encompasses α-fucosidases,

as well as family GH20, which includes enzymes with hexosaminidase and lacto-*N*-biosidase activities (*38*). An example of host-produced glycans metabolised by bifidobacteria are human milk oligosaccharides (HMOs), which are present in human breast milk, but are not utilized by the infant host. Genomic analysis of a typical faecal isolate from breast-fed infants, *B. longum* subsp. *infantis (B. infantis)*, has revealed the presence of a 43-kb gene cluster predicted to encode GHs and carbohydrate transporters necessary for the import and metabolism of HMOs, including α-fucosidases, sialidases and β-galactosidases, as well as secreted solute binding proteins and permeases (*33, 59, 60*)

The evaluation of the bifidobacterial GH repertoire also involved the clustering of bifidobacterial species and subspecies based on their predicted GHs and carbohydrate degradation pathways (*38*). This analysis has allowed the identification of three distinct groups, which the authors designated GHP/A, GHP/B, and GHP/C. Group GHP/A has been found to encompass bifidobacterial species and subspecies with a considerable range of predicted GH43 family members. These enzymes are involved in the breakdown of complex plant glycans such as xylan and arabinoxylans, which represent substantial components of plant cell wall-derived dietary fibres. This suggests that members of the group GHP/A have evolved adaptations to hosts with vegetarian or omnivorous diet (*38*). Members of the second group, GHP/B, have been shown to have a limited number of enzymes belonging to families GH43 and GH3, with members of the latter family involved in the bacterial cell wall biosynthesis and hydrolysis of such sugars as cellodextrin, (arabino)xylan and (arabino)galactan. This indicates adaption to an omnivorous host for members of this group (*38*). Finally, bifidobacterial taxa isolated from social insects clustered to form the GHP/C group and have been found to have a distinct set of enzymes belonging to GH43 and GH3 family, but a limited repertoire of GH13 family.

Part of the predicted bifidobacterial glycobiome has been found to be extracellular and involved in the degradation of polysaccharide polymers too large to be internalized. Approximately 11% of the total GH repertoire has been estimated to be located extracellularly. The extracellular GHs encompass members of the GH13

family (annotated as pullulanases and α-amylases), members of the GH43 family (β-xylosidases and α-L-arabinofuranosidases), and members of the GH51 family (α-L-arabinofuranosidases). Predicted secreted GHs have been found to be present in 43 out of 47 recognised species and subspecies of the genus *Bifidobacterium*. The genomes of *B. biavatii, Bifidobacterium scardovii*, and *B. bifidum* have been found to encompass the highest number of extracellular enzymes. Interestingly, the genome of *B. bifidum* has been predicted to include members of GH83 and GH33, which are putative sialidases. This finding has suggested advanced genetic adaptation of *B. bifidum* to the mammalian gut, as sialidases are crucial for the degradation of HMOs and intestinal glycoproteins, such as mucin (*61, 62*). However, *B. bifidum* has been shown to be unable to use sialic acid as its sole carbon source thus the activity of sialidases has been suggested to provide access to other carbohydrates associated with sialylated host glycans (*63*). In addition, it has been proposed that sialic acid released as a result of sialidase activity can be utilised by other bifidobacteria, for example *B. breve*, in cross-feeding interactions (*63, 64*).

1.6 *Bifidobacterium* as members of the wider gut microbiota of humans

Infancy lasts from birth to approximately two years of age and is characterised by rapid growth, development and maturation of organs and systems (*65*). It has been well recognised that the numbers of bifidobacteria in humans decrease over lifetime, however the observations on bifidobacterial abundance and species diversity during infancy have not been consistent (*66*). Findings from a number of studies have indicated that members of *Bifidobacterium* dominate the infant gut microbiota (*67, 68*), however, other reports have shown fluctuations in numbers of these bacteria, including very low abundance or even absence in particular individuals (*66, 69, 70*). These inconsistent results could potentially be explained by variation in cohort age, size, geographical location or methodology between studies, including differences in experimental protocols, sequencing technology and data analysis. For example, the accuracy of studies based on the identification of the 16S rRNA gene strongly depends on the choice of primers and the balance

between efficiency, specificity and sensitivity in the targeting of the different bacterial 16S rRNA gene sequences in samples (*71*). Overall, *B. longum* subsp. *infantis*, *B. longum* subsp. *longum* and *B. bifidum* have been indicated to dominate the infant microbiota, while *B. adolescentis* and *B. longum* subsp. *longum* have mainly been associated with the "adult" gut (*72*). The results of a longitudinal study that followed Japanese infants during the first three years of life showed that at the age of 3, *B. longum* and *B. breve* followed by *Bifidobacterium catenulatum* and *B. bifidum* constituted the predominant *Bifidobacterium* species in the microbiota (*66*). Another investigation of healthy Japanese adults revealed that the bifidobacterial community was dominated by *B. longum*, *B. catenulatum* and *B. adolescentis (73)*. These findings suggested that at the age of 3, the *Bifidobacterium* population is in transition and stabilises later in life (*66*).

1.7 Diet and early life development

According to the concept of "nutritional programming", first introduced in the early 1990s, the quality and the quantity of nutrients consumed by infants during the first year of life can permanently affect, or "programme" the early-life developmental outcomes (*74*). It has been proposed that exposure to specific stimuli or insults during infancy, considered a critically important period of development, can exert long-lasting effects on the host.

The scientific evidence gathered over the years linking nutrition in early life to health in adulthood, was the basis for health promotion and the establishment of nutrition programmes around the world. In their 2009 report, the British Medical Association Board of Science (*75*) recognised the importance of infant nutrition and its association with the lifelong health. The 2011 report of the UK Scientific Advisory Committee on Nutrition (*76*) recommended the optimisation of the diet and the body composition of young women, and the promotion and support of breastfeeding, based on the compelling evidence for a role of early-life nutrition in modulating risk of such medical conditions as coronary heart disease, type-2 diabetes, osteoporosis, asthma, lung disease and certain types of cancer.

More recently, additional evidence has been provided linking balanced co-maturation of the gut microbiota and the host with infant wellness and development, resulting in lifelong health benefits. It is now well established that diet, along with factors such as genetics, mode of birth and antibiotic use, plays a significant role in shaping bacterial communities across different body sites. Crucially, establishment of the early life microbiota within the infant gastrointestinal tract drives key aspects of host development, occurring alongside host growth and immune development (*77-80*). Notably, recent studies have associated the composition of the gut microbiota with the risk of developing a number of chronic diseases, including type-2 diabetes (*81, 82*) cardiovascular disease (*83*), and cancer (*84, 85*).

1.8 Diet as a factor modulating the gut microbiota development – an overview

Diet has been proposed to be one of the major factors modulating the development of human microbiota. The type of feeding (breast vs formula) and the time of the introduction of solid foods (*86*) were suggested to be crucial to maturation of the early life microbial communities. The infant gut microbiota is highly unstable until the age of 2-3, when it reaches composition similar to that of an adult. Compared to infants, adults harbour a "hardier" microbiota, which is not as prone to external factors that might cause large changes in its structure, however, recent studies have shown that changes in diet can lead to temporary microbial shifts within hours even in adults (*87*). These finding are in line with the replacement of milk-based diet with one that is more diverse and based on solid foods during weaning (*86*).

Transition from breastfeeding to a more complex diet, coupled with the introduction of complementary foods is considered to lead to an overall increase in the bacterial diversity in the infant gut, with the establishment of Firmicutes and Bacteroidetes as dominant phyla (*88-90*). Thompson et al. (*91*) compared the composition of the microbiota between infants that were exclusively breast-fed and

those that were non-exclusively breast-fed during weaning and found differences in bacterial diversity between these two groups. Species such as *Veillonella*, *Roseburia*, and members of the *Lachnospiraceae* family were found present in exclusively breast-fed infants following the start of complementary feeding, while *Streptococcus* and *Coprobacillus* were identified in non-exclusively breast-fed infants after solid foods were introduced. Interestingly, an increase in the relative abundance of *Bifidobacterium* during complementary feeding was observed in non-exclusively breast-fed infants, while the relative abundance of this genus in the exclusively breast-fed infants showed a decrease. These results may indicate differential modulatory effects of starches and other complex carbohydrates introduced in diet during weaning (*91*).

Several studies have reported shifts in the composition of the infant microbiota during the complementary feeding period in late infancy (+9 months). In line with previous studies (*70, 92*), Laursen et al. (*90*) have shown an increase in the relative abundance of families *Lachnospiraceae, Ruminococcaceae, Eubacteriaceae, Rikenellaceae, Sutterellaceae*, and a decrease in *Bifidobacteriaceae, Actinomycetaceae, Veillonellaceae, Enterobacteriaceae, Lactobacillaceae, Enterococcaceae, Clostridiales incertae sedis XI*, *Carnobacteriaceae*, and F*usobacteriaceae* in infants over time. The overall increase in the bacterial diversity during weaning has been linked to functional changes in the microbiome. The results of the analysis of predicted metabolic pathways between infants that were exclusively and non-exclusively breast-fed indicated a higher abundance of genes encoding sugar transporters in breast-fed infants compared to those that were not exclusively breast-fed, which can possibly be linked to the abundance of HMOs in the diet of breast-fed infants (*91*). In contrast, infants that were non-exclusively breast-fed showed an enrichment in genes involved in nitrogen and methane metabolism, as well as peptidases, which could potentially be attributed to the higher protein content in their diet (*91*).

The transition from milk-based diet to solid foods has been considered to initiate the development of a more stable, and functionally more complex, adult-like microbiome harbouring genes involved in the degradation of complex

carbohydrates, starches, and xenobiotics, as well as vitamin production (*88*). Several studies reported increased total levels of SCFAs during weaning, in particular butyrate and propionate, which has been associated with an increase in abundance of Firmicutes and Bacteroidetes (*88, 92*). Widely reported increase in the abundance of members of family *Lachnospiraceae* and a decrease in saccharolytic *Bifidobacterium* associated with the introduction of complementary feeding have been correlated with the increased protein content of the diet, while higher numbers of *Prevotellaceae* have been linked to the presence of fibre in complementary foods (*93*).

1.9 Prebiotics and their role in optimising early life nutrition

The manipulation of the intestinal microbiota for health improvement is currently being investigated worldwide. The introduction of specific dietary components i.e. prebiotics, with the potential to preferentially "feed" and act upon beneficial members of the gut microbiota is one widely used approach. Gibson and Roberfroid (*94*) first introduced the concept of prebiotics, and defined them as "non-digestible food ingredients that beneficially affect the host by selectively stimulating the growth and/or activity of one or a limited number of bacteria in the colon, thus improving host health". In 2010, the International Scientific Association for Prebiotics and Probiotics (ISAPP) updated this definition to incorporate the functional aspect of these substances and described them as "selectively fermented ingredient that results in specific changes in the composition and/or activity of the gastrointestinal microbiome, thus conferring benefit(s) upon host health" (*95*). Bindels et al. (*96*) have suggested an even more comprehensive definition: "a non-digestible compound that, through its metabolization by microorganisms in the gut, modulates composition and/or activity of the gut microbiome, thus conferring a beneficial physiologic effect on the host". The newest definition of prebiotics accommodates scientific considerations and reflects its importance for regulatory bodies, industry, and consumers worldwide (*97, 98*). According to Bindels et al. (*96*), all well-known prebiotics are complex carbohydrates. Other nutrients, for example bioactive secondary plant metabolites such as polyphenols, may also have

prebiotic properties. Since prebiotics are stable compounds, they can easily be added to different types of food, such as yogurt, ice cream, biscuits, breads, cereals, spreads and drinks (*99*). To be classified as a prebiotic, a compound needs to fulfil the following criteria (i) it needs to resist gastric acidity, degradation by mammalian digestive enzymes, and gastrointestinal absorption; (ii) it has to be metabolised by intestinal bacteria; and (iii) it needs to selectively stimulate the growth and/or activity of intestinal bacteria linked with health and well-being (*100*).

To date, a number of prebiotics originating from different sources and displaying various chemical properties have been named, reviewed and classified based on a set of common criteria (*101*). Examples will be discussed in more detail below, but briefly; properties of breast milk-derived HMOs as first prebiotics available to infants have been widely recognised (*102-104*). Other well-established compounds with prebiotic properties include inulin, fructo-oligosaccharides (FOS), galacto-oligosaccharides (GOS), and lactulose, while isomalto-oligosaccharides (IMO), xylo-oligosaccharides (XOS), and lactitol are categorised as emerging prebiotics (*101, 105-108*). Inulin-derived FOS from chicory root, as well as wheat bran-derived XOS and arabinoxylo-oligosaccharides (AXOS) have been shown to have applications in human and animal welfare (*109, 110*). Mannitol, maltodextrin, raffinose, lactulose, and sorbitol have been identified to exert beneficial effects on members of the human gut microbiota, including promotion of growth and production of antimicrobial compounds (*111-113*). In addition, resistant starches and other dietary fibres, such as beta-glucan and guar gum, can be degraded by intestinal bacteria to SCFAs, which indicates potential prebiotic application of these compounds in promoting human well-being (*114-116*).

1.10 Breast milk: gold standard infant nutrition and a source of beneficial microbes

Human milk is perfectly adapted to neonatal nutritional requirements and supports infant growth and development. Hence, it has been considered the gold standard for infant nutrition during the first months of life (*117, 118*). Breastfeeding drives beneficial microbial, metabolic and immunological programming, and thus

represents a key postnatal link between mother and infant (*119-121*). A large number of biologically active components are present in human milk, including immunoglobulins, chemokines, growth factors, cytokines, bioactive lipids, oligosaccharides, microRNAs, hormones, immune cells and microorganisms (*120, 122-124*). Crucially, breastfeeding has been linked to decreased risk for numerous diseases, which has been attributed to differences in immune system development (*124, 125*). Biologically active factors in human milk have been shown to directly impact immune responses (*124, 126*). Milk glycoproteins have been reported to prevent intestinal pathogens, such as *Vibrio cholerae* and *Escherichia coli*, from colonising (*127*). Human milk oligosaccharides have been shown to prevent attachment of respiratory pathogens, such as *Streptococcus pneumoniae*, to respiratory epithelium (*128*). In particular, lacto-*N*-neotetraose and lacto-*N*-tetraose have been proposed as receptors for attaching pneumococci, with the disaccharide GIcNAcβ1→3Galβ identified as the principal binding site (*129*). HMOs have also been found to directly interact with glycan-binding proteins expressed on the epithelial cells and cells of the innate immune system (*130*). For example, 2'-fucosyllactose (2'-FL) has been reported to bind human DC-SIGN (Dendritic cell-specific intercellular adhesion molecule-3-grabbing non-integrin), a C-type lectin receptor present on the surface of macrophages and dendritic cells (*131*). Breast milk components have also been found to facilitate the establishment of early bacterial colonisers which further modulate immunity and enhance the synthesis and secretion of polymeric IgA, the coating antibody that protects mucosal surfaces against bacterial invasion, and helps drive homeostatic immune cell programming including the balance between Th1 vs. Th2 responses. Notably, unbalanced immunological T helper cell responses (Th1 > Th2 or Th2 > Th1) have been associated with certain clinical outcomes, such as atopic disease (Th2 imbalance) and Crohn's disease (Th1 imbalance) (*126*).

The dogma that breast milk is sterile was challenged after viable bacteria were isolated from breast milk samples obtained from healthy women (*132, 133*). Traditional culture methods have reported the presence of lactic acid bacteria with potential beneficial or "probiotic" properties such as *Lactobacillus, Lactococcus,*

Leuconostoc, and also in some cases *Bifidobacterium*, as well as other bacteria including *Streptococcus, Enterococcus* and *Staphylococcus* (*132-135*). Studies involving sequencing have indicated high microbial diversity, and detected bacteria that typically inhabit the oral cavity, such as *Veillonella* and *Prevotella*, skin bacteria such as *Staphylococcus* and *Propionibacterium*, lactic acid bacteria, including *Enterococcus*, *Streptococcus, Leuconostoc* and *Weissella*, and Gram negative bacteria, such as *Pseudomonas, Ralstonia,* and *Klebsiella*, with high inter- and intra-individual specific profiles (*134, 136-141*). Breast milk, the maternal gut and infant faeces and oral samples have been reported to share specific bacteria or bacterial DNA (*133, 142-145*), including specific strains belonging to *Bifidobacterium* and *Staphylococcus* spp. (*142, 146-148*). These findings support the notion of vertical transmission of breast milk bacteria to the infant gut, whose roles likely include enhancement of the development of the early life immune system, maintenance of tolerance to commensal microbial members, and intestinal host defence to pathogens. Currently there is limited evidence with regard to the potential effect of breast milk microorganisms on infant health, but it has been speculated that they may act as key early life colonisers, and help to shape the developing mucosal immune system (*123, 149, 150*). More specifically, the breast milk bacterial communities may promote intestinal immune homeostasis, encouraging a shift from the predominant intrauterine T helper cell 2 (Th2) immune milieu, to a Th1/Th2 balanced response, and to trigger regulatory T cell differentiation (*123, 151*).

1.11 Breast milk "feeds" specific members of the infant gut microbiota

Maternal breast milk acts as a natural prebiotic and feeds certain bacterial strains and species of the gut microbiota during the early life developmental window. Particular breast-milk components, e.g. HMOs, are poorly digested by the infant but are suitable growth substrates mainly for *Bifidobacterium*, but also a few strains of *Bacteroides* and *Lactobacillus* that can enzymatically degrade these complex dietary components. The term "human milk oligosaccharides" collectively refers to non-digestible carbohydrates with the degree of polymerisation (DP) equal or above 3,

that are the third most abundant component of breast milk after lactose and lipids (*152*). More than 200 different HMOs can be found in the milk of a single mother, each composed of a common lactose reduced end linked to *N*-acetyllactosamine (Galβ1-4GlcNAc) or lacto-*N*-biose I units (LNB) (Galβ1-3GlcNAc) (*153*). Several HMOs comprise the major types in breast-milk, for example 2'-FL and lacto-*N*-tetraose (LNT). These structures can additionally be sialylated or/and fucosylated. HMO fucosylation depends on the presence of functional enzymes, fucosyltransferases, in the mother's mammary glands, which is regulated by Lewis and secretor genes (*153, 154*).

Predominance of *Bifidobacterium* in the gut microbiota during infancy and lower overall bacterial diversity have been associated with nutrient availability. The high abundance of *Bifidobacterium* in breast-fed infants has been linked to the enrichment of members of this genus in genes required for the degradation of HMOs (*70, 155*). Several infant gut-associated *Bifidobacterium* species have been found to harbour genes encoding enzymes necessary for the metabolism of HMOs, including *B. breve, B. bifidum, B. longum* subsp. *longum (B. longum), B. longum* subsp. *infantis (B. infantis),* and more rarely *Bifidobacterium pseudocatenulatum* (*59, 156-158*). Strategies for HMO degradation vary among *Bifidobacterium* species, and the metabolic capabilities for HMOs within the species are strain specific. Overall, conserved consumption of HMOs has been attributed to *B. bifidum* and *B. infantis*, while more variable and strain-dependent degradation of these carbohydrates has been demonstrated in *B. breve* and *B. longum* (*152, 159*).

Recent experimental evidence has suggested that bifidobacteria utilise two main strategies for HMO metabolism (*152, 160*). The first one relies on presence of functional membrane-associated extracellular glycosyl hydrolases, lacto-*N*-biosidases. These enzymes hydrolyse HMOs, particularly LNT, into mono- and disaccharides, which are then imported into the cell by solute binding proteins (SBP) (*159*). This strategy seems to be widely employed by *B. bifidum*, but also by those *B. longum* strains that harbour the gene encoding lacto-*N*-biosidase (LnbX) (*152, 160*). The second strategy relies on the SBP-mediated import of intact HMOs inside the cell and their degradation by intracellular glycosyl hydrolases, including

for example α-fucosidases, sialidases, β-galactosidases and β-*N*-acetylhexosaminidases with affinities for a wide spectrum of milk oligosaccharides (*159*). This mode of action has been demonstrated in *B. infantis*, *B. breve* and those *B. longum* strains that are LnbX-negative (*152, 160*).

Certain *Bifidobacterium* members, for example *B. bifidum*, seem to only partially degrade HMOs, which raised a question of possible cross-feeding within infant-associated *Bifidobacterium* communities (*161*). Gotoh et al. (*152*) found that wild-type strain of *B. longum* 105-A (LnbX-positive), which showed very limited growth on LNT in mono-culture, grew very well in co-culture with an extracellular HMO degrader *B. bifidum*. In addition, supplementation of *B. bifidum* in HMO-containing stool cultures stimulated the growth of other *Bifidobacterium* species, especially in samples which did not initially contain *B. bifidum*. Most recently, our own results suggested similar co-operation between HMO degraders that employ the intracellular degradation system and other members of *Bifidobacterium* community (*162*). We demonstrated that conditioned media from *B. longum* grown on 2'-FL used as a carbon source for other non-HMO degrading *Bifidobacterium* strains could support growth of other isolates within the same infant host. These results were further confirmed in co-culture experiments. Altogether, these results suggest that bifidobacterial degradation products resulting from HMO metabolism may influence the acquisition and abundance of *Bifidobacterium* in an ecosystem (*152, 162*).

Previous studies have indicated that members of genus *Bifidobacterium* strongly modulate specific immune cells and pathways. Acquisition of specific strains and species of *Bifidobacterium* seem to correlate with a critical period of immune maturation and programming. It has been shown that the metabolism of maternal-derived glycans, i.e. HMOs, by *Bifidobacterium* results in increased levels of SCFAs, in particular acetate (*162-164*). These compounds have been reported to play regulatory roles within the gut-associated mucosal immune system. They have also been found to provide energy to epithelial cells, which in turn helps maintain the coherence of tight junctions and the intestinal barrier, thus providing mechanisms of defence against pathogens. SCFAs have also been implicated in regulating

mucosal cells of the innate immune system such as macrophages or dendritic cells, as well as neutrophils, and the antigen-triggered adaptive immunity activities mediated by T and B lymphocytes (*165-168*).

1.12 The formula effect: impact on the infant gut microbiota

Factors such as mode of birth (i.e. vaginal delivery vs. Caesarean-section), antibiotic treatment, and skin-to-skin contact determine the composition of early life microbiota and the resulting gastrointestinal function (*93, 169*). However, the infant diet is proposed to play a major role and induce the most significant changes. Over the years, numerous studies have reported notable differences in the composition of intestinal microbiota between breast- and formula-fed infants (*92, 170-174*). Compared to breast-fed infants, formula-fed infants are characterised by an overall higher microbial diversity, elevated levels of potentially pathogenic organisms, such as *Escherichia coli* and *Clostridioides difficile* and reduced levels of beneficial *Bifidobacterium* (*70, 175, 176*).

Formula feeding has been linked to increased risk for a number of diseases in adulthood, attributed to microbiota profiles and the previously mentioned development of the immune system in early life (*125*). Numerous epidemiological studies have suggested that the type of feeding in infancy may contribute to obesity or the development of type-2 diabetes later in life (*177-180*). Some studies have also suggested an increased risk of high blood pressure (*181*) and high cholesterol (*182, 183*).

Over the years, improved formulations of infant formulas have been researched, designed and produced as an alternative to breast milk. These products must contain water, carbohydrates, protein, fat, vitamins and minerals in adequate proportions (*184*). The composition of infant formulas is tightly regulated, and manufacturers around the world are obliged to follow guidelines set by government agencies (*184*). All major components (protein, lipids, carbohydrates) added to the formula must have a proven history of safe use (*185*). Fructose should be avoided as a carbohydrate source, as it may lead to fructose intolerance in infants, and only

the L forms of amino acids are approved as ingredients, as D forms may cause D-lactic acidosis (*186*). Further, hydrogenated fats are also not allowed, nor is the use of the ionizing radiation on the product (*187*).

Currently available infant formulas can be grouped in three major classes: cow-milk based formula, soy-based formula and specialized formula. They differ in nutritional and caloric values, taste, digestion, and cost. Special formulas have been designed and are available to meet needs of infants with intolerances, where rice, amino acids and extensively hydrolysed whey or casein proteins have been used as cow's milk substitutes (*184*). HMOs are characteristic of human milk and are not present in the same composition in animal milk nor plant proteins that are normally used as the basis for formula (*188*). Synthesising HMOs is expensive, therefore formula supplementation with compounds that might mimic some of the HMO functions described above has been widely investigated in the recent years. Affordable non-digestible carbohydrates (NDC) that have been used for this purpose include galacto-oligosaccharides (GOS), fructo-oligosaccharides (FOS) and pectin oligomers (POS) (*189*).

1.13 Optimisation of infant formulas with pre- and probiotics

Galacto-oligosaccharides (GOS): are produced by microbial glycosyl hydrolases by the transglycosylation of lactose and are comprised of galactosyl residues ranging from 2 to 10 units, and a terminal glucose linked by β-glycosidic bonds, whose type is determined by the source of enzyme (*190*). The β-glycosidic linkages are resistant to the action of gastrointestinal enzymes, which allows GOS to reach the colon and undergo degradation by bacteria (*191*).

Positive effects of adding GOS to the formula on the numbers of bifidobacteria and lactobacilli have been shown previously. Ben et al. (*192*) and Fanaro et al. (*193*) have reported that feeding formula supplemented with GOS to infants increases the numbers of *Bifidobacterium* spp. after 3 months, compared to infants fed unsupplemented formula. Using an *in vitro* model of the proximal colon,

Maathuis et al. (*194*) showed that representatives of genera *Bifidobacterium* and *Lactobacillus*, namely *B. longum, B. bifidum, B. catenulatum, Lactobacillus gasseri*, and *Lactobacillus salivarius* directly degraded GOS. Other bacteria, including members of Bacteroidetes and Firmicutes could also ferment these carbohydrates, but to a lesser degree (*194*). In another study, Watson et al. (*195*) have demonstrated that particular strains of bifidobacteria and lactobacilli can selectively metabolise GOS *in vitro*. Sierra et al. (*196*) conducted a clinical trial, whose results have indicated a significant increase in the proportion of acetic acid and the decrease in butyric acid in infants fed with formula supplemented with GOS. This pattern is similar to that described in breast-fed infants and can be correlated with higher abundance of *Bifidobacterium*. Further, the same study has shown that a significantly higher proportion of infants fed the GOS-containing formula were colonised with *B. adolescentis* at 4 months of age (*196*).

Fructans: are polymers, whose structure consists of a linear chain of fructose units with a terminal glucose. The chains are linked by β-glycosidic bonds, which make these carbohydrates resistant to digestive enzymes (*190*). Chains vary in length between specific fructans and can contain up to 200 residues. Fructans are heterogeneous and have been organised into subclasses based on the length of chain; short-chain fructo-oligosaccharides (scFOS) with 3 to 5 residues per chain, fructo-oligosaccharides (FOS) containing 6 to 10 residues, and structurally similar inulin and levan with a DP ranging from below 10 to 200 residues (*190*). Short-chain FOS and FOS are produced by transfructosylation of sucrose from sugar beet. They can also be directly extracted from such plants as onions, garlic, wheat and bananas. Additionally, fructans can be commercially produced by enzymatic digestion of long-chain inulin (*190*).

Bifidobacteria and lactobacilli have been shown to selectively metabolise scFOS and fructo-oligosaccharides *in vitro*. Marx et al. (*197*) have demonstrated that particular strains of *B. adolescentis, B. longum, B. breve* and *B. pseudocatenulatum* are capable of metabolising FOS. Interestingly, *B. adolescentis* demonstrated the best growth, the highest degree of acidification and was the only strain capable of degrading both short- and long-chain FOS. Growth of other strains either slowed or

stopped once short-chain fructans were consumed. Rossi and co-workers (*198*) screened a significant number of human- and animal-associated *Bifidobacterium* strains for growth on FOS (n=55), and concluded that while all strains fermented fructo-oligosaccharides, most of them could not degrade inulin. In line with previous studies, tested *Bifidobacterium* strains showed preference for scFOS as a substrate. Similarly, Kaplan and Hutkins (*199*) have assessed the ability of different strains of lactobacilli to ferment scFOS, and found that the majority of tested strains were able to ferment scFOS to SCFAs.

Studies looking at the effects of fructan supplementation on the composition of the gut microbiota in both infants and adults have associated fermentation of FOS, or mixtures of FOS and GOS, with increased numbers of beneficial bacteria, namely genera *Bifidobacterium* and *Lactobacillus*, and a decrease in abundance of potentially pathogenic bacteria, such as *Clostridium perfringens* and *E. coli* (*200-203*). These results have been explained by the increased production of SCFAs as a result of fructan degradation. SCFAs lower pH in the gut, which in turn inhibits the growth of a number of pH-sensitive pathogens, members of *Enterobacteriaceae* and Clostridia (*204, 205*). Furthermore, Paineau et al.(*206*) found that infants who were fed formula containing scFOS had increased abundance of bifidobacteria at 2 and 3 months of age compared to the placebo group.

Pectins: are structurally complex polysaccharides with backbone mainly composed of galacturonic acid linked by α−1,4 glycosidic bonds. The backbone, as well as side chains, can incorporate other monosaccharides, including arabinose, rhamnose and galactose (*207*). It has been estimated that over 67 transferases are involved in the biosynthesis of pectins, due to the complexity of these carbohydrates. Pectins are classified according to the degree of methyl esterification, which affects the plant cell wall structure and properties (*207*).

Di et al. (*208*) evaluated the potential prebiotic effects of five pectic oligosaccharides (POS) produced by enzymatic treatment and acid hydrolysis of citrus peel. The results indicated that POS had bifidogenic effect in *in vitro* batch fermentation experiment, resulting in the increase in the numbers of *Bifidobacterium* over time. The authors indicated that the prebiotic activity of POS

was structure-dependent and linked to arabinose-rich oligosaccharides. Several species of *Bacteroides*, as well as the Firmicutes *Eubacterium eligens* and *Faecalibacterium prausnitzii* have also been shown to degrade pectins (*209, 210*). Recently, it has been suggested that prebiotic formulations containing pectin-derived acidic oligosaccharides have the ability to mimic the effects of the acidic fraction of HMOs and may have similar effects to those of HMOs in terms of immune modulation and the development of immune responses in infants (*190*). Stam et al. (*211*) have demonstrated that a formulation comprised of GOS, long-chain FOS, and acidic pectic oligosaccharides had no adverse effects on vaccine specific antibody levels in healthy term infants. Furthermore, authors postulated that this formulation could promote Th1- and Treg-dependent immune responses and induce a downregulation of IgE allergic responses. The results from animal studies were in line with these observations and demonstrated that a mixture of GOS, FOS and arabino-oligosaccharides gave an optimal Th1-dependent delayed-type hypersensitivity (DTH) response and reduced Th2 cytokine production in young adult mice, compared to a GOS/FOS-containing formulation (*212*). Another study suggested that the same combination of oligosaccharides decreased symptoms of allergic asthma in mice (*213*). Some pectins have been suggested to exert undesired effects in *in vitro* models (*190*). The work of Bosscher et al. (*214*) indicated that the supplementation of infant formulas with either esterified pectin or locust bean gum resulted in decreased availability of calcium, iron, and zinc compared to standard infant formula and human milk. These results may suggest a decreased bioavailability of these nutrients *in vivo*.

Human Milk Oligosaccharides (HMOs): Although supplementation of infant formulas with NDC has been a cost-effective method to provide health benefits associated with complex carbohydrates to infants not consuming maternal breast-milk, prebiotic plant-derived oligosaccharides cannot substitute all the beneficial functions of HMOs (*191*). Globally, much effort has been directed towards production of HMOs on a commercial scale. Chemical synthesis is one approach that could deliver well-defined carbohydrate structures, however reliable methods for large-scale production are not available (*191*). Instead, enzymatic and

chemoenzymatic HMO synthesis methods have been researched. A number of studies successfully accomplished production of HMOs *in vitro* using bacterial transferases (*215, 216*). The use of engineered bacteria for HMO production has been approved by the U.S. Food and Drug Administration (FDA) and led to the commercialisation of large-scale production of one of the major HMOs, 2'-FL (*191*). A number of expression systems have been designed and made available in *E. coli* and *Saccharomyces cerevisiae* (*217-219*). In addition, bacterial glycosyl hydrolases, widely used to produce GOS on an industrial scale, have been researched as an alternative for the production of HMOs (*220, 221*). The transglycosylation activity of these enzyme can be exploited to attach carbohydrate moieties to existing sugars, resulting in the formation of fucosylated or sialylated glycans mimicking HMOs (*222*). This approach could potentially be employed to manufacture rare complex polysaccharides that are currently commercially unavailable (*191*).

Several companies have implemented these innovative methods to produce synthetic HMOs to date. In 2016, Glycom A/S (DK) received authorisation from the European Commission to introduce two of their products obtained through fermentation by engineered bacteria, 2'-*O*-fucosyllactose and lacto-*N*-neotetraose (LNnT), as new food ingredients on the European market (*223, 224*). This was followed in 2017 by an authorisation for sialic acid manufactured by enzymatic synthesis to be used in infant nutrition, follow-on formulas, and baby foods (*225*). Similarly, Jennewein Biotechnology GmbH (DE) and FrieslandCampina (NL) received the Novel Food authorization for 2'-FL (*226*). This technology has already been implemented in infant formulas. Abbot was the first company to launch a product (Similac Pro-Advance) containing one of these HMOs, 2'-FL, in the US in 2016. Nestlé followed with Nan Optipro Supreme containing both 2'-FL and LNnT, which was launched in Spain, formulas containing 2'-FL introduced in Asia under the Illuma brand, and SMA Nutrition products containing 2-FL and LNnT available in the UK and Ireland.

Probiotics: have been defined by the World Health Organisation (WHO) and the Food and Agricultural Organisation of the United Nations (FAO) as "live microorganisms (bacteria or yeasts), which when ingested or locally applied in

sufficient numbers confer one or more specified demonstrated health benefits for the host" (*227*). Following recent scientific and clinical developments, the International Scientific Association for Probiotics and Prebiotics (ISAPP) have reviewed and updated this definition, with a minor grammatical correction, to "live microorganisms that, when administered in adequate amounts, confer a health benefit on the host" (*228*). The consensus statement published in 2014 has proposed a set of benchmark standards for the differentiation of probiotic products and clarified the proper scope and appropriate use of the term probiotic (*228*). Probiotic bacteria typically belong to the genera *Lactobacillus* and *Bifidobacterium* (*229, 230*). Health benefits associated with these organisms resulted in their widespread use as ingredients in functional foods, including a broad range of dairy and non-dairy products (*231, 232*).

Probiotic supplementation of infant formulas has been explored as a method to replicate the health benefits associated with breastfeeding (and the transfer of breast milk microbial communities) and to establish an intestinal microbiota similar to that of breast-fed infants in those infants that are not consuming mother's milk (*169, 233*). In their position paper, the European Society for Paediatric Gastroenterology, Hepatology and Nutrition (ESPGHAN) has recommended that infant formulas with added probiotics should not be marketed unless their health benefits and safety have been fully documented (*234*). Production of infant formulas falls under strict regulations globally, and the probiotics aimed to be used as supplements in infant foods are subject to a range of regulatory requirements (*169*). In Europe, the European Food Safety Authority (EFSA) has introduced a premarket system for safety assessment of microorganisms used in food and feed (*235*), while in the U.S., the FDA has implemented Evidence-Based Review System for the Scientific Evaluation of Health Claims to provide guidelines for producers (*236*).

Since beneficial effects of probiotics are strain specific, manufacturers are obliged to scientifically prove probiotic properties and cell viability when making a health claim related to a specific product (*169, 233, 234, 237*). According to Kent and Doherty (*169*), four probiotic strains whose functions have been documented in

numerous scientific studies, namely *Bifidobacterium animalis* subsp. *lactis* (*B. lactis*) BB-12, *Lactobacillus rhamnosus* GG, *Lactobacillus acidophilus* NCFM (along with BB-12) and *Lactobacillus reuteri* DSM 17938, are routinely used by large formula producers, such as Nestle and Heinz. Furthermore, dried probiotic formulations containing cocktails of lactobacilli and bifidobacteria that can be added to breast-milk, human milk fortifiers and formulas have been made commercially available in recent years, however their health claims often prove difficult to validate due to the lack of strain designations and scientific evidence for health benefits (*169, 238*).

Concerns regarding the safety of administering probiotics to infants have been raised in terms of their immunogenic effects and their impact on infant growth, as well as potential bacterial infection due to the underdeveloped immune system (*169, 239*). The widespread use of probiotic formulations in infants requires careful and comprehensive surveillance to monitor for unintended consequences. Van den Nieuwboer at al. (*240*) conducted a systematic analysis of 57 clinical trials and eight follow-up studies, and concluded that the use of probiotics in infants between zero and two years of age is safe with regard to the evaluated strains in infants with a particular health status or susceptibility. However, long-term effects of probiotic administration to infants have rarely been measured. In addition, questions remain about the ability to identify infants at risk at the time of birth (*239*).

1.14 Life after milk: the influence of additional complex dietary components on the early life gut microbiota during the crucial weaning window

The composition of complementary foods plays a major role in the transition from simple to adult-like microbiota during weaning, providing essential nutrients to the infant and the developing intestinal microbial communities while stimulating the immune development (*241*). Starchy foods constitute a common group of complementary foods due to their texture and palatability (*241*). Plant-derived complex carbohydrates in complementary foods have been proposed to exert beneficial health effects on infants through their bifidogenic and prebiotic properties and the resulting modulation the intestinal microbiota and metabolic

end products (*55, 242*). Examples of such non-digestible carbohydrates with prebiotic effects include poly- and oligosaccharides composed of fructose and a terminal glucose (e.g. inulin-type fructans (ITF)), or arabinose and xylose (e.g. arabinoxylans (AX) or arabinoxylo-oligosaccharides (AXOS)) (*243, 244*).

Inulin: is widely distributed in nature and can be found in fruit and plants such as chicory roots, onions, bananas and leeks (*243*). Structurally, inulin is composed of a backbone of fructose monomers linked by β-2,1 bonds with a DP between 2 and 65, often linked to a terminal glucose monomer by an α-1,2 bond. Oligofructose is a subgroup of inulin and consists of polymers with a DP below 10 (*245*). β-fructofuranosidases, which cleave terminal fructose residues from the non-reducing ends of the fructose polymers, have been shown to act on inulin (*246*). Several of these enzymes have been characterised in several species of *Bifidobacterium*, including *B. infantis* (*247*), *B. lactis* (*248*) and *B. adolescentis* (*249*).

The ability of *Bifidobacterium* to degrade inulin are species- and strain-specific. Falony et al. (*250*) identified four different clusters of strains representing ten different *Bifidobacterium* species based on their inulin degradation capabilities. Cluster A could only metabolise fructose, while cluster B was able to degrade both fructose and oligofructose after importing the latter inside the cell. Clusters C and D were found to degrade both inulin and oligofructose extracellularly, followed by the release of fructose into the medium. Selak et al. (*251*) examined the ITF fermentation capabilities of 190 strains of *Bifidobacterium*, representing five species and originating from different regions of the intestine, and discovered that the degradation properties of different strains were not correlated to the region in the intestine. Interestingly, strains with different ITF metabolism capabilities were found coexisting in the same intestinal region, which suggested bacterial cooperation in the degradation of ITF, and a potential for cross-feeding interactions between strains or at the species level (*251*). Several previous reports indicated cross-feeding activities between bifidobacteria and butyrate producing bacteria grown simultaneously on ITF, where the consumption of ITF by bifidobacteria resulted in production of acetate, which was in turn used as a co-substrate to produce butyrate by colonic butyrate producers *Roseburia intestinalis* DSM 14610

(*252*), *Roseburia inulinivorans* DSM 16841^T (*253*), and *Faecalibacterium prausnitzii* DSM 17677^T (*254*). Proposed effects of the consumption of ITF on human health include increased absorption of calcium and magnesium in the colon, increased secretion of satiety hormones, increased stool frequency and a decrease in proteolytic fermentation in distant colon (*255*).

Arabinoxylan (AX) and arabinoxylo-oligosaccharides (AXOS): have also received scientific attention in recent years due to their prebiotic and bifidogenic properties. Arabinoxylan (AX) is a hemicellulose consisting of a linear backbone of β-1,4 xylose residues with arabinose substitution, found in all major cereal grains, with the highest content found in rye followed by wheat, barley, oats, rice, and sorghum (*256*). The cleavage of the internal β-xylosidic linkages of AX by endo-1,4-β-xylanases to xylo-oligosaccharides results in production of arabinose-substituted xylo-oligosaccharides (AXOS) (*257*). Complete degradation of complex structures of AX and AXOS requires a cooperation of a number of enzymes with debranching and depolymerising actions, including afore-mentioned endo-β-xylanases, as well as β-xylosidases, exo-oligoxylanases and α-arabinofuranosidases (*55*).

Bioinformatic analyses have identified genes encoding enzymes involved in the cleavage of the AX and AXOS backbones in genomic sequences of several bifidobacterial strains, some of which have been expressed and characterised (*258*). Examples include α-arabinofuranosidases in *B. adolescentis* ATCC 15703^T, *B. adolescentis* DSM 20083, *B. longum* B667, and *B. longum* NCC2705; β-xylosidases in *B. adolescentis* ATCC 15703^T and *B. lactis* BB-12; and exo-oligoxylanases in *B. adolescentis* LMG 10502 (*259-261*). According to Riviere et al. (*55*) no endo-β-xylanases have been characterised in *Bifidobacterium*, so it is likely that members of this genus have to cooperate with other bacterial species, for example *Roseburia* or *Bacteroides*, to achieve complete degradation of AX and AXOS.

The bifidogenic properties of AX and AXOS have been reported in a number of studies. Hughes et al. (*262*) demonstrated *in vitro* that fermentation of wheat-derived AX resulted in proliferation of members of the genera *Bifidobacterium*, *Lactobacillus* and *Eubacterium*. Following a clinical study, Chung et al. found that in

addition to an increase in *Bifidobacterium* during AXOS supplementation, the proportional abundance of *Prevotella* species increased significantly when these organisms were present in the baseline microbiota (*263*). In an *in vivo* study with rats, Van Craeyveld et al. (*264*) showed that the bifidogenic effect was only caused by AXOS with low average DPs ≤ 5 and A/X ≤ 0.27, while another study reported a 60-fold increase in *Bifidobacterium* in the cecum of rats fed long-chain AX with average DP = 60 and A/X of 0.70 (*265.*). The findings from the latter study were further validated by another study which reported the presence of two different *B. longum* species during the fermentation of long-chain AX in an *in vitro* model of the proximal colon (*266*). Riviere et al. (*267*) demonstrated that the ability of *Bifidobacterium* to metabolise AX and AXOS is strain-dependent. The authors identified five bacterial clusters capable of differential degradation of AXOS. Cluster I contained fifteen strains, representatives of 7 different bifidobacterial species (*B. adolescentis, Bifidobacterium angulatum, B. bifidum, B. breve, Bifidobacterium dentium, B. longum,* and *Bifidobacterium thermophilum*), which were not able to metabolise neither arabinose substitutions nor the xylan backbone of AXOS, although some strains could utilise arabinose and xylose. Members of the identified Cluster II, eight *B. longum* strains, could not utilise the xylan backbone of AXOS, but were able to degrade the arabinose substitutions. Cluster III was comprised of ten strains spanning six bifidobacterial species (*B. adolescentis, B. angulatum, B. longum, B. animalis, Bifidobacterium gallicum,* and *B. pseudolongum*), which were capable of utilising the xylan backbone of AXOS up to xylotetraose but could not metabolise the arabinose substitutions or showed limited capabilities to do so. Cluster IV contained only two strains of *B. longum*, which shared the ability to degrade AXOS, while Cluster V was comprised of a single strain of *B. catenulatum* LMG 11043, which could use all XOS fractions in a non-preferential way and displayed broad degradation capabilities for arabinose substitutions (*267*).

Being slowly-fermentable carbohydrates, AX and AXOS have been associated with a number of health benefits. These polysaccharides have been suggested to affect the production of toxic metabolites by protein- and lipid-fermenting microbes in the distal colon through stimulation of saccharolytic bacterial species, such as

Bifidobacterium, which results in increased production of SCFAs leading to the lowering of the luminal pH (*264, 268, 269*). In obese mice, AXOS consumption and the resulting increase in *Bifidobacterium* has been proposed to help restore gut barrier functions and cure endotoxemia (*270*). Furthermore, consumption of AX and AXOS has been linked to increased colonic absorption of dietary minerals, increased antioxidant capacity, reduced post-prandial glycaemic response and reduced blood cholesterol (*244, 271-273*).

Resistant starches (RS): are homo-polysaccharides of glucose that escape the digestion in the upper gastrointestinal track (*274*). They have been categorised into four main types. RS1 is physically indigestible, for example starches in whole grains. RS2 is a granular starch that is present in such foods as green bananas, uncooked potato or high amylose maize. RS3 are retrograded starches produced by cooking and cooling processes, whereas RS4 are starches that have been chemically modified by esterification, crosslinking, or transglycosylation and are not found in nature (*274, 275*).

Modulatory effects of resistant starches on the host gut microbiota have previously been reported by a number of *in vivo* studies. Tachon et al. (*276*) demonstrated that mice fed diets containing RS2 type starch from high-amylose maize were colonized by higher numbers of Bacteroidetes and species of *Bifidobacterium*, *Akkermansia*, and *Allobaculum* in proportions dependent on the starch concentration. For example, mice fed 18% resistant starch were mostly colonized by close relatives of *B. pseudolongum* subsp. *pseudolongum* and *globosum*, whereas those fed 36% RS were enriched with *Bifidobacterium* most closely related to *B. animalis*. Additionally, the levels of *Bifidobacterium* and *Akkermansia* were associated with expression levels of proglucagon, the precursor of the gut anti-obesity/diabetic hormone GLP-1, mice feeding responses and the weight of the gut (*276*). The prebiotic effects of resistant starches were also demonstrated in adult humans. Martinez et al. (*277*) reported varying effects of starch types on the composition of the human faecal microbiota. Type RS4 was associated with changes at the phylum level, significantly increasing Actinobacteria and Bacteroidetes while decreasing Firmicutes. In addition, at the species level, consumption of RS4 was

correlated with an increase in *B. adolescentis* and *Parabacteroides distasonis*, while RS2 and RS3 significantly raised the numbers of *Ruminococcus bromii* and *Eubacterium rectale* (*277, 278*).

Abilities of *Bifidobacterium* to ferment starch and starch derivatives were tested *in vitro* by Duranti et al. (*279*) using representatives of 47 known bifidobacterial taxa. The results indicated that *B. adolescentis, B. angulatum, B. boum, B. breve, B. cuniculi, B. dentium, B. longum, B. infantis, Bifidobacterium merycicum, Bifidobacterium minimum, Bifidobacterium reuteri, B. pseudocatenulatum, Bifidobacterium ruminantium,* and *B. thermophilum* grew substantially (OD > 0.5) on RS2 resistant starch as sole carbon source. The study identified two strains which exhibited very good growth among representatives of human-associated *B. adolescentis*, out of which *B. adolescentis* 22L, isolated from human milk, could grow on starch and all starch derivatives (maltodextrin, maltotriose, pullulan, and glycogen). Genomic and transcriptomic analyses of *B. adolescentis* 22L revealed the presence of a number of glycosyl hydrolases putatively implicated in degradation of starch, as well as starch-derived glycans (*279*). They included four proteins with predicted α-amylase activity, a glycogen/starch phosphorylase, putative phosphoglucomutase and 4-α-glucanotransferase, as well as two genes predicted to encode amylopullulanases, similar to the extracellular starch-degrading enzyme previously characterised in *B. breve* UCC2003 (*280*). In addition, genomic analyses identified two genes predicted to encode α-1,4- and α-1,6-glucosidases, an ABC-type system, and a gene encoding a LacI-type regulator flanking the gene encoding an amylopullulanase, whose products may also be involved in starch utilization using mechanisms similar to those described for *B. breve* UCC2003 (*279, 280*). More recently, Jung et al. (*281*) have also identified genes encoding enzymes implicated in direct or indirect degradation of RS and its derivatives in another *Bifidobacterium* strain, *Bifidobacterium choerinum* FMB-1 isolated from rumen fluids of cattle. The putative enzymes included α-amylase, maltosyltransferase, pullulanase type I, glycogen phosphorylase, glycogen debranching enzyme, 4-α-glucanotransferase, and glucan branching enzyme. Interestingly, the above-mentioned study by Duranti et al. (*279*) did not identify the representative type strain of *B. choerinum* LMG

10510^T as one that could utilise starch, which could further suggest that RS utilisation in *Bifidobacterium* is strain-dependent.

Maintenance of healthy blood glucose levels has been widely proposed as an important health benefit of starch fermentation by a number of studies (*282*) (*283, 284*). Bacterial fermentation of RS and associated changes in the gut microbiota composition have also been associated with the increase in the production of SCFAs, and a reduction in secondary bile acids, phenol and ammonia (*277, 285*).

Although the above refers to human hosts, these dietary components are also present in the diets of different animal hosts and will be described in more detail in Chapters 4 and 5.

1.15 Bifidobacterium as members of the animal gut microbiota

In recent years, the use of next-generation sequencing has significantly facilitated the investigations of the bifidobacterial diversity and contributed to the discovery of novel *Bifidobacterium* species. The majority of described *Bifidobacterium* type strains have been isolated from animal hosts, however sufficient genomic information for the non-human-associated taxa is lacking (*31, 286, 287*). Human-derived strains, in particular those associated with infancy and belonging to *B. longum*, *B. bifidum*, *B. breve and B. pseudocatenulatum* species, have received much attention in recent years due to their potential probiotic properties, leading to considerable isolation and characterisation efforts (*152, 159, 162, 288-291*). As a result, data on human-associated bifidobacteria are currently overrepresented in databases. For example, out of a total of around 2,100 *Bifidobacterium* genomes deposited at the NCBI Genome database (July 2020) (*5*), more than 400 sequences are available for *B. longum*, more than 100 for both *B. breve* and *B. bifidum*, and more than 90 for *B. pseudocatenulatum*.

With regard to animal-associated *Bifidobacterium*, only two species, namely *B. animalis* (both subsp. *animalis* and subsp. *lactis*) and *B. pseudolongum* (both subsp. *pseudolongum* and subsp. *globosum*), have been more widely studied to date (*292, 293*). The isolation and sequencing efforts resulted in the availability of

considerable amount of data on these species, with more than 90 sequences available for *B. animalis* and more than 70 genomes available for *B. pseudolongum* (NCBI Genome, July 2020). Recent genomic investigations into these data confirmed that both *B. animalis* and *B. pseudolongum* are widely distributed among animal hosts belonging to different animal classes, e.g. mammals and birds, and thus appear to be generalist in terms of host-specificity (*292, 293*). In addition, experimental approaches indicated subspecies-specific differences in carbohydrate metabolism capabilities within *B. animalis* and *B. pseudolongum* species and suggested that these differences may result from evolutionary adaptations of members of particular subspecies to their respective environmental niches. For example, strains of *B. animalis* subsp. *animalis*, isolated primarily from rodents, were shown to metabolise a broad range of carbohydrates, including arabinose, galactose, glucose, maltose, melibiose, sucrose, and xylose, whereas those identified as *B. animalis* subsp. *lactis*, recovered mainly from humans and primates, seemed to be more specialised in the fermentation of lactose, maltose, raffinose, and sucrose. Similarly, strains of *B. pseudolongum* subsp. *globosum* isolated from ruminants displayed higher growth performances on carbohydrates such as cellobiose, rhamnose, starch and trehalose, with limited growth on glucose and lactose. The latter sugars are metabolised in the rumen and are scarce in the ruminant large intestine. In contrast, *B. pseudolongum* subsp. *pseudolongum* strains isolated from pigs were shown to grow significantly better on such carbohydrates as glucose, glycogen, lactose, maltodextrin, maltose and melibiose (*292, 293*).

Widespread *Bifidobacterium* distribution across mammals have also been suggested for species other than *B. animalis* and *B. pseudolongum*, based on the ITS profiling of the 45 type strains of the genus *Bifidobacterium* across 291 faecal samples collected from 67 host species spanning 10 mammalian taxonomic orders (*294*). Sequences corresponding to *B. longum* and *B. adolescentis* were shown to be ubiquitously distributed in the profiled mammalian species, with a prevalence of 95.5% and 91%, respectively, followed by *B. pseudolongum* and *B. bifidum* with a prevalence of 85%. In addition, sequences corresponding to species traditionally

associated with particular groups of hosts, for example *Bifidobacterium actinocoloniiforme*, *Bifidobacterium asteroides* or *Bifidobacterium indicum* associated with insects, were shown to be present among mammals belonging to different taxonomic groups. In contrast, sequences corresponding to the type strains of more recently characterised *Bifidobacterium* species isolated from non-human primates, namely *Bifidobacterium aesculapii* and *Bifidobacterium tissieri*, were associated with hosts belonging to the order Primates. Moreover, this analysis revealed the presence of 89 sequences with an identity level of <93% to the 45 analysed bifidobacterial species in the faecal samples of mammals, suggesting that a considerable number of putative novel *Bifidobacterium* species and subspecies remain to be discovered (*294*).

Indeed, the number of recognised *Bifidobacterium* species has significantly increased in recent years, from 47 in October 2016 (start of my PhD) to 87 in July 2020, with the majority of newly characterised species recovered from non-human primates. A recent analysis of the distribution of recognised primate-associated *Bifidobacterium* in faecal samples collected from 57 human and non-human primates revealed a further potential for the discovery of novel species in this group of animals based on the 16S rRNA profiling, with the average relative abundance of putative new species in human and monkey faecal samples estimated at 6% and 28%, respectively (*295*). The members of the primate family *Cebidae* have been predicted to harbour the highest number of predicted novel *Bifidobacterium* species, with an average percentage of putative novel taxa estimated at 46%. In addition, the analysis of co-phylogenetic relationships between 24 primates and 23 bifidobacterial species suggested the existence of strong co-phylogenetic patterns between human-associated *Bifidobacterium* (*B. adolescentis, B. bifidum, B. breve, B. catenulatum, B. dentium, B. longum spp.* and *B. pseudocatenulatum*) and the members of *Hominidae* family (*Gorilla gorilla*, *Homo sapiens* and *Pan troglodytes*). Furthermore, *B. adolescentis*, *B. biavatii* and *B. ramosum* were indicated as the taxa most commonly shared among primates (*295*).

Despite an increase in the availability of genomic data, robust studies focusing on the diversity and the functional properties of members of *Bifidobacterium* in animal

hosts remain lacking, with the pan-genome and the phylogenetic relationship of the genus unresolved. The majority of previous studies investigating the diversity of *Bifidobacterium* have either focused on a subset of human-associated isolates or examined single type strains as representatives of species, and in consequence included a limited number of the isolates from different habitats (*27, 38, 93, 159, 162, 286, 289*). The sampling among host animals is quite uneven, and *Bifidobacterium* species isolated from certain host groups are significantly underrepresented in the databases (e.g. those from insects and reptiles). Further isolation efforts and genomic investigations into the diversity of *Bifidobacterium* are crucial for the comprehensive understanding of the distribution of members of this genus across the animal kingdom, the mechanisms behind their beneficial properties, and their role as members of the wider gut microbiota.

1.16 Hypotheses

1.16.1 Overarching hypothesis

Bacteria belonging to the genus *Bifidobacterium* are important members of the gut microbiota of humans and animals, with specific strains displaying beneficial properties associated with their capabilities to ferment complex carbohydrates.

1.16.2 Study specific hypotheses

Chapter 3: Whole genome sequences of *Bifidobacterium longum* display particular genomic features related to carbohydrate metabolism. These features may play a role in the establishment and persistence of members of this species in individual hosts during infancy despite dietary changes associated with transition from breast milk to solid foods.

Chapter 4: Whole genome sequences of *Bifidobacterium* isolated from small mammals display particular genomic features related to carbohydrate metabolism, host modulation and defence mechanisms.

Chapter 5: The diversity of *Bifidobacterium* in animal hosts is not fully explored. Experimental and computational methods (bacterial isolation and genomic analysis, respectively) contribute to the discovery of true diversity within the genus. Whole genome sequences of animal-associated isolates display particular traits related to their isolation source and functional capabilities, e.g. degradation of carbohydrates.

Chapter 2

Materials and methods

2.1 Materials

2.1.1 Equipment and reagents

Primary reagents, kits, equipment and tools used or mentioned in this chapter are listed below:

Equipment	Model	Supplier
Anaerobic chamber	Ruskinn Concept Plus	Baker Ruskinn, Bridgend, UK
Autoclave (portable)	Classic 2100 Standard (9-litre)	Prestige Medical, Blackburn, UK
Centrifuge	Eppendorf 5810R	Eppendorf, Stevenage, UK
Homogeniser for DNA extraction samples	FastPrep-24	MP Biomedicals, Santa Ana, USA
Microcentrifuge	Prism	Labnet, New Jersey, USA
Microplate reader	Tecan infinite F50	Tecan, Männedorf, Switzerland
PCR station	UV HEPA PCR Systems	UVP, Upland, USA
PCR thermal cycler	Applied Biosystems Veriti 96- well thermal cycler	Thermo Fisher Scientific, Loughborough, UK
pH meter	Martini MI 151	Rocky Mount, USA
Qubit meter	Qubit 2.0	Thermo Fisher Scientific, Loughborough, UK
Safety cabinet	Walker Class II MSC	Walker Safety Cabinets, Glossop, UK
Shaking Incubator	Stuart	Cole-Palmer, Illinois, USA

Table 2.1 Primary equipment used in laboratory

Item	Supplier
Agar	Oxoid (Thermo Fisher Scientific), Loughborough, UK
Brain Heart Infusion (BHI)	Oxoid (Thermo Fisher Scientific), Loughborough, UK
DNA lo-bind tubes (1.5 and 2.0ml)	Eppendorf, Stevenage, UK
FastDNASPIN kit for Soil	MP Biomedicals, Santa Ana, USA
Inoculation sterile loop (disposable)	Microspec, Wirral, UK
KAPA 2G Robust PCR kit	KAPA Biosystems, Wilmington, USA
Qubit dsDNA BR assay	Invitrogen (Thermo Fisher Scientific), Loughborough, UK
Reinforced Clostridial Medium (RCM)	Oxoid (Thermo Fisher Scientific), Loughborough, UK
Spreader (disposable)	Microspec, Wirral, UK
BD Difco™ Lactobacilli MRS Broth	BD Biosciences, New Jersey, USA
Sodium iododacetate	Oxoid (Thermo Fisher Scientific), Loughborough, UK
Mupirocin	PanReac AppliChem (VWR), Pennsylvania, USA
Cysteine HCl	Sigma-Aldrich, Dorset, UK

Table 2.2 Primary materials and kits used in experiments

2.1.2 Faecal samples, bacterial isolates and isolate DNA extracts

Samples (faecal samples, bacterial isolates and DNA) were obtained through collaboration with the following organisations:

Collaborator	Organisation	Sample cohort	Sample type	Location
Dr Anne McCartney	University of Reading	Breast- and formula-fed infants Breast- and formula-fed infants	DNA from bacterial isolates Bacterial isolates	Reading, UK
Dr Sarah Knowles	University of Oxford	Wild small mammals	Faecal samples	Oxford, UK
Dr Laima Baltrūnaitė	Nature Research Centre	Wild small mammals	Faecal samples	Vilnius, Lithuania
Sara Goatcher	Banham Zoo and Africa alive	Captive animals	Faecal samples	Suffolk, UK

Table 2.3 Sample sources for particular projects

2.1.2.1 Faecal sample collection for breast- and formula-fed infant study (*B. longum*) (Chapter 3)

Faecal sample collection was performed by Dr Anne McCartney. Infants were recruited between 2005 and 2007: five were exclusively breast-fed and four were exclusively formula-fed. Faecal samples were obtained from infants at specific intervals during the first 18 months of life. For inclusion in the study, infants had to meet the following criteria: have been born at full-term (>37 weeks gestation); be of normal birth weight (>2.5 kg); be <5 weeks old and generally healthy; and be exclusively breast-fed or exclusively formula-fed [SMA Gold or SMA White (Wyeth Pharmaceuticals), to avoid supplemented formulae and to keep consistency within the formula group]. The mothers of the breast-fed infants had not consumed any antibiotics within the 3 months prior to the study and had not taken any prebiotics and/or probiotics. Ethical approval was obtained from the University of Reading Ethics Committee (*296*). For details, refer to paper by Rogers and McCartney from 2010 (*296*). All faecal samples were processed by members of Dr Anne McCartney's research team; for the purpose of this project I was provided with bacterial isolates and DNA samples from the isolates.

2.1.2.2 Faecal sample collection for wild mammal study (Chapter 4)

Rodent trapping and faecal sample collection were performed by Dr Sarah Knowles, Miss Aura Raulo and Dr Laima Baltrūnaitė. In the UK, live rodent trapping was carried out at two sites of mixed deciduous woodland approximately 50km apart: Wytham Woods (51° 46'N, 1°20'W) and Nash's Copse, Silwood Park (51°24'N, 0°38'W). All animals were live-trapped using a standard protocol across both sites, using small Sherman traps baited with peanuts and apple and provisioned with bedding, set at dusk and collected at dawn the following day. All newly captured individuals were marked with subcutaneous PIT-tags for permanent identification, and all captures were weighed and various morphometric measurements taken. Faecal samples were collected using sterilised tweezers from the base of Sherman traps into sterile tubes. Samples from Silwood were frozen within 8 hours of

collection at -80°C and sent frozen to the Quadram Institute (Norwich, UK) for *Bifidobacterium* culturing; samples from Wytham were posted on the day of sampling at room temperature to the same address for culturing. Upon reception of samples, they were immediately frozen at -80°C. To ensure no cross-contamination and identification of samples to specific individuals, any traps that showed signs of rodent presence (captures and trigger failures) were washed thoroughly in a bleach solution and autoclaved between uses.

In Lithuania, small mammals were trapped in October 2017 and between May - November in 2018 using live- and snap-traps at twelve locations: Site 1: 54.92878, 25.33333; Site 2: 55.02814, 25.27380; Site 3: 54.92866, 25.31524, Site 4: 55.05938, 25.35643, Site 5: 54.96362, 25.35640; Site 6: 54.93026, 25.24138; Site 7: 54.99276, 25.24909, Site 8: 55.02700, 25.35867; Site 9: 55.06756, 25.29782; Site 10: 54.76482, 25.31283; Site 11: 54.76322, 25.35052; Site 12: 54.93650, 25.28442. Traps baited with bread soaked in sunflower oil (in case of live-traps, apple and bedding were also added) were set in the evening and retrieved in the morning. Small mammals trapped with snap-traps were placed in separate bags and transported to the lab on ice. Live-trapped animals were transported to the lab and humanely killed by cervical dislocation. Species, sex, age, reproduction status of small mammals were identified. Content of distal part of colon (~20-30 mm) was removed, placed in Eppendorf tube and stored at –80°C. Frozen samples were sent to the Quadram Institute (Norwich, UK) for *Bifidobacterium* culturing.

The distance between trapping sites in both the UK and Lithuania was far enough for the animals not to move between them on a regular basis. All studied species have small home ranges – *Apodemus* spp. have the widest range and rarely move more than 0.25 km (*297*).

2.1.2.3 Faecal sample collection for captive animal study (Chapter 5)

Samples were collected by animal-care staff at either Banham Zoo and Africa Alive! into sterile Sterilin specimen containers with spoon and stored under anaerobic

conditions using Oxoid AnaeroGen 2.5L sachets at 4 °C. Samples were transported to the laboratory and stored at -80 °C within 48 hours.

2.1.3 Media and bacterial isolation

2.1.3.1 Testing of alternative agar media

Depending on the amount of available faecal material, samples (~50mg or ~100mg) were re-suspended in either 450µl or 900µl of sterile Phosphate Buffer Saline and used to produce serial dilutions (neat to 10^{-4}). The samples were vortexed for 30s and mixed on a shaker at 1600rpm. The dilutions were plated on media composed of either MRS broth, BHI (Brain Heart Infusion) broth or RCM broth with addition of bacteriological agar (15g/l), antibiotic mupirocin (50mg/l) and L-cysteine hydrochloride monohydrate (50mg/l). Furthermore, all three abovementioned agar plate types were prepared with the addition of sodium iodoacetate (25mg/l) to test whether the addition of this reagent to the media would increase their selectivity. Sodium iodoacetate inhibits the glycolytic enzyme glyceraldehyde-3-phosphate dehydrogenase (GAPDH), whose activity is not an obligate requirement for carbohydrate metabolism in heterolactic bacteria, including *Bifidobacterium* (*298, 299*). The plates were incubated for 48 to 72 hours at 37^{o}C in an anaerobic cabinet, after which single colonies with morphology consistent with that of bifidobacteria were re-streaked with disposable sterile culture loops onto fresh agar plates and incubated. To obtain pure colonies, the re-streaking process was repeated 3 times.

2.1.3.2 Bacterial isolation – wild mammal study (Chapter 4)

Faecal samples were prepared as above and plated onto BHI agar supplemented with mupirocin (50mg/l), L-cysteine hydrochloride monohydrate (50mg/l) and sodium iodoacetate (7.5mg/l) and incubated in an anaerobic cabinet for 48-72 hours. Three colonies from each dilution were randomly selected and streaked to purity on BHI agar supplemented with L-cysteine hydrochloride monohydrate (50mg/l). Pure cultures were stored in cryogenic tubes at -80^{o}C.

2.1.3.3 Bacterial isolation – captive animal study (Chapter 5)

Faecal samples were prepared as above and plated onto either BHI agar supplemented with mupirocin (50mg/l), L-cysteine hydrochloride monohydrate (50mg/l) and sodium iodoacetate (7.5mg/l) or MRS agar supplemented with mupirocin (50mg/l) and L-cysteine hydrochloride monohydrate (50mg/l). Plates were incubated in an anaerobic cabinet for 48-72 hours. Three colonies from each dilution were randomly selected and streaked to purity on either BHI agar or MRS agar supplemented with L-cysteine hydrochloride monohydrate (50mg/l). Pure cultures were stored in cryogenic tubes at -80^{o}C.

2.1.4 Bacterial cultures

Pure colonies were grown in 20ml of MRS broth in sterile 50ml centrifuge tubes to obtain numbers of bacteria sufficient for further procedures.

2.1.5 Bacterial stocks

Stocks were made in cryovials using RCM with glycerol (30% V/V) and stored at -80^{o}C.

2.2 DNA extraction

2.2.1 FastDNA^{TM} SPIN kit method

This method was employed for preliminary strain identification based on partial 16S rRNA gene sequences. Bacterial pellets were re-suspended in 980µl of sodium phosphate buffer, transferred to Lysing Martix E tube and 120µl of MT buffer was added to the suspension. Bacteria were then homogenised for 3min at speed 6.0, using FastPrep® 24 instrument, and centrifuged at 14000g for 15min. 250µl of PPS was added to a clean microcentrifuge tube, sample supernatant was transferred to this, after which the tubes were mixed by hand 10 times and centrifuged at 14000g for 10 min. 1ml of Binding Matrix was added to clean 15ml tubes, and sample

supernatant was transferred to this. The samples were inverted for 2 min and placed on a rack for 3 min to allow the Binding Matrix to settle. 500µl of the supernatant was removed and discarded, after which the Binding Matrix was re-suspended in the remaining supernatant. 700µl of the mixture was transferred to SPIN filter and centrifuged at 14000g for 2min, after which catch tubes were emptied. This procedure was repeated until all the Binding Matrix mixture was used. The pellet in the SPIN filter was re-suspended in 500µl of SEWS-M, and centrifuged at 14000g for 5 min. Catch tubes were then emptied and samples were centrifuged a second time (dry spin) for 5 min at 14000g. Catch tubes were discarded, replaced with clean Eppendorf LoBind microcentrifuge tubes, and the samples were air dried for 10min at room temperature. 65µl of DES was then added to the samples, and the samples were incubated at room temperature for 5min. Finally, the samples were centrifuged at 6600g for 2min to bring eluted DNA into Eppendorf tubes. Extracted DNA was stored at 4°C until required for PCR.

2.2.2 Phenol-chloroform method

This method was used to extract high-quality bacterial genomic DNA from strains identified as *Bifidobacterium* based on the 16S rRNA gene. Bacterial pellets were re-suspended in 2ml of 25% sucrose in 10mM Tris and 1mM EDTA at pH 8.0. Cells were then treated using 50µl of 100mg/ml lysozyme. Further, 100µl of 20mg/ml Proteinase K, 30µl of 10mg/ml RNase A, 400µl of 0.5 M EDTA (pH 8.0) and 250µl of 10% Sarkosyl NL30 were added into the lysed bacterial suspension. The samples were then incubated on ice for 2 hours, followed by 50°C overnight water bath.

Next, samples were subject to three rounds of Phenol:Chloroform:Isoamyl Alcohol (25:24:1) extraction using Qiagen MaXtract High Density tubes. Further two rounds of extractions with Chloroform:Isoamyl Alcohol (24:1) were then performed to remove residual phenol, followed by ethanol precipitation and 70% ethanol wash, after which DNA pellets were resuspended in 300µl of 10mM Tris (pH8.0). Sample DNA concentration was quantified using Qubit dsDNA BR Assay Kit in Qubit 2.0 Fluorometer according to the manufacturer's protocol. Extracted DNA was stored in -20°C until further analysis.

Phenol-chloroform DNA extraction for the breast- and formula-fed infant study (*B. longum*) (Chapter 3) was carried out by Dr Anne McCartney.

Item	Components and conditions	Supplier
25% sucrose in TE buffer	TE buffer: 10mM Tris and 1mM EDTA pH8.0	Sigma-Aldrich, Dorset, UK
Chloroform:IAA	24:1 Chloroform:IAA	Sigma-Aldrich, Dorset, UK
E buffer	10mM Tris pH8.0	Sigma-Aldrich, Dorset, UK
EDTA	0.5M EDTA, pH8.0 (pH adjusted)	Sigma-Aldrich, Dorset, UK
Ethanol	Absolute ethanol (>99.5%) MaXtract High Density tubes	Sigma-Aldrich, Dorset, UK
Lysozyme	Lysozyme re-suspended in 0.25M Tris pH8.0	Roche, West Sussex, UK
Phase-lock gel tubes	MaXtract High Density	Qiagen, Manchester, UK
Phenol-Chloroform-Isoamyl alcohol (IAA) solution	25:24:1 (Phenol-Chloroform-Isoamyl Alcohol)	Sigma-Aldrich, Dorset, UK
Phosphate Buffer Saline (PBS)	PBS, pH7.2	Sigma-Aldrich, Dorset, UK
Proteinase K	~15-20mg/ml proteinase K	Roche, West Sussex, UK
RNase A	10mg/ml RNase A (boiled before use)	Roche, West Sussex, UK
Sarkosyl NL30	30% Sarkosyl	Sigma-Aldrich, Dorset, UK

Table 2.4 Materials used in phenol-chloroform DNA extraction for whole genome sequencing.

2.3 DNA Sequencing

2.3.1 Sequencing of the 16S rRNA gene for preliminary bacterial identification

2.3.1.1 PCR, primers, conditions

Concentration of extracted DNA was adjusted for PCR to 10-20ng/µl using molecular H_2O. Kapa2G Robust PCR reagents and molecular H_2O were used to prepare the PCR master mix (10mM of dNTP mix (1µl/sample), Kapa2G GC buffer with $MgCl_2$ (10µl/sample), molecular H_2O (31.6µl/sample), Kapa2G Robust – Taq polymerase (0.4µl/sample), with addition of 10 µM primers (1µl/sample; Table 2.5) (*300*).

Primers	Sequence
fD1	5'- AGA GTT TGA TCC TGG CTC AG - 3'
fD2	5'- AGA GTT TGA TCA TGG CTC AG - 3'
rP1	5'- ACG GTT ACC TTG TTA CGA CTT - 3'

Table 2.5 Primers used for PCR amplification of 16S rDNA.
Primer abbreviations: f - forward; r - reverse; D - distal; P - proximal. Reverse primers produce sequences complimentary to the rRNA.

5µl of diluted bacterial DNA was mixed with 45µl of the master mix and amplified using Veriti™ 96 well thermal cycler as follows: polymerase activation at 94°C for 5min, followed by 35 cycles of denaturation at 94°C for 1min, primer annealing at

43°C for 1min, strand extension at 72°C for 2min and final strand extension at 72°C for 7min.

2.3.1.2 16S rRNA gene sequencing

Un-purified PCR products were prepared according to sample submission guide for Value Read sequencing service (Sanger sequencing, Eurofins, Luxembourg). Most final 16S rRNA gene product sizes ranged between 900-1000bp.

2.3.1.3 Whole genome sequencing

This work was performed at the Wellcome Trust Sanger Institute (Hinxton, UK) and at the Quadram Institute Bioscience (Norwich, UK).

At Hinxton, DNA libraries were prepared as detailed in Table 2.6 and Table 2.7 below and subjected to WGS pipeline on Illumina HiSeq 2500 platform to generate 125bp paired-end reads.

Step	Description	Details
1	Quantitation of DNA	Biotium Accuclear Ultra high sensitivity dsDNA Quantitative kit
2	DNA shearing	Mechanical shearing using Covaris LE220 instrument
3	Purification of sheared samples	Agencourt AMPure XP SPRI beads on Agilent Bravo WS
4	Library construction (end-repair, A-tailing and ligation)	'NEB Ultra II custom kit' on an Agilent Bravo WS automation system
5	PCR	KapaHiFi Hot start mix and IDT 96 iPCR tag barcodes on Agilent Bravo WS automation system
6	Post-PCR purification	Agencourt AMPure XP SPRI beads on Beckman BioMek NX96 liquid handling platform
7	Quantitation of libraries	Biotium Accuclear Ultra high sensitivity dsDNA Quantitative kit
8	Pooling libraries	Pool in equimolar amounts on a Beckman BioMek NX-8 liquid handling platform
9	Normalisation	Libraries normalised to 2.8nM
10	Loading to sequencer	Loading on requested Illumina sequencing platform

Table 2.6 Library preparation for whole genome sequencing (Illumina HiSeq 2500).

PCR Step	Temperature	Duration	Cycles
Initial denaturation	95	5 min	1
Denaturation	98	30 s	5
Annealing	65	30 s	
Extension	72	1 min	
Final extension	72	10 min	1

Table 2.7 PCR conditions for WGS libraries (Illumina HiSeq 2500).

At Quadram Institute Bioscience, genomic DNA was normalised to 0.5ng/µl with EB (10mM Tris-HCl). 0.9µl of TD Tagment DNA Buffer (Illumina Catalogue No. 15027866) was mixed with 0.09µl TDE1, Tagment DNA Enzyme (Illumina Catalogue

No. 15027865) and 2.01µl PCR grade water in a master mix and 3µl added to a chilled 96 well plate. 2µl of normalised DNA (1ng total) was pipette mixed with the 3µl of the tagmentation mix and heated to 55 ^{o}C for 10 minutes in a PCR block. A PCR master mix was made up using 4µl kapa2G buffer, 0.4µl dNTPs, 0.08µl Polymerase and 6.52µl PCR grade water, contained in the Kap2G Robust PCR kit (Sigma Catalogue No. KK5005) per sample and 11µl added to each well need to be used in a 96-well plate. 2µl of each P7 and P5 of Nextera XT Index Kit v2 index primers (Illumina Catalogue No. FC-131-2001 to 2004) were added to each well. Finally, the 5µl of Tagmentation mix was added and mixed. The PCR was run with 72^{o}C for 3 minutes, 95^{o}C for 1 minute, 14 cycles of 95^{o}C for 10s, 55^{o}C for 20s and 72^{o}C for 3 minutes. Following the PCR reaction, the libraries were quantified using the Quant-iT dsDNA Assay Kit, high sensitivity kit (Catalogue No. 10164582) and run on a FLUOstar Optima plate reader. Libraries were pooled following quantification in equal quantities. The final pool was double-SPRI size selected between 0.5 and 0.7X bead volumes using KAPA Pure Beads (Roche Catalogue No. 07983298001). The final pool was quantified on a Qubit 3.0 instrument and run on a High Sensitivity D1000 ScreenTape (Agilent Catalogue No. 5067-5579) using the Agilent Tapestation 4200 to calculate the final library pool molarity. The pool was run at a final concentration of 10pM on an Illumina MiSeq instrument.

2.4 Bioinformatics

2.4.1 Computing environment and resources

Norwich Bioscience Institutes (NBI) High-Performing Computing Cluster (HPC) with CentOS Linux operating system and MacBook Pro with MacOS High Sierra updated to MacOS Catalina v10.15.4 were used to run command line bioinformatic tools. On HPC, all software and dependencies were loaded on to the shell environment before usage. SLURM workload manager (*301*) (v16.05.8) was used to submit computational jobs to HPC.

2.4.2 Preliminary 16S rRNA gene sequence analysis

For primary bacterial strain identification, the 16S rRNA gene sequencing results were compared with NCBI 16S ribosomal RNA sequences database (Bacteria and Archaea) using the Nucleotide BLAST algorithm (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Classification was assigned based on highest identity hits and similarity scores (≥97% threshold was used as an indicator that the isolate belongs to the genus *Bifidobacterium*).

2.4.3 Genome assembly and annotation

Bacterial genomes were either assembled at Wellcome Trust Sanger Institute using Velvet v1.2.10 (*302*) or in-house using SPAdes v3.11 (*303*). Velvet assemblies were submitted to screen using Kraken v1.1 (MiniKraken) (*304*) to check for contamination at 5% threshold. Genome assemblies found to be contaminated with >5% of other bacterial species were excluded from analyses. Contigs below 500bp were filtered out of the assemblies.

The remaining sequencing reads, generated on both Illumina 2500 and MiSeq instruments, were screened for contamination using Kraken v1.1 (MiniKraken) at 5% threshold and pre-processed with fastp v0.20 with default settings (*305*). SPAdes v3.11 with “careful” option was used to produce assemblies, after which contigs below 500bp were filtered out. All genomes were annotated with Prokka v1.13 (*306*).

2.4.4 Publicly available genomes

Lists of publicly available genomes downloaded from NCBI and used in this work are included in supplementary material associated with respective thesis chapters: Table S3.2, Table S4.1 and Table S5.2. Where used for comparative analysis, downloaded genomes were annotated with Prokka v1.13 for processing consistency.

2.4.5 Average Nucleotide Identity calculation

Python3 module pyANI v0.2.7 (breast- and formula-fed infant study) updated to v0.2.10 (wild mammal study and captive animal study) with default BLASTN+ settings was employed to calculate the average nucleotide identity (ANI) (*307*). Species delineation cut-off was set at 95% identity (*308*). Genomes displaying ANI value >99.9% were considered identical (*309*).

2.4.6 Phylogenetic analysis of strain LH_867 - wild mammal study (Chapter 4)

As experimental 16S rRNA gene amplification only yielded a partial gene sequence (987bp), the location of 16S rRNA genes was predicted in whole genome sequence of strain LH_867 as well as the genomes of type strains using barrnap v0.9 (*310*) and extracted *in silico*. In the case of *Bifidobacterium simiarum* the 16S rRNA gene sequence could not have been predicted, and as it is not available for download from public databases, this strain was excluded from the 16S rRNA gene analysis. The predicted 16S rRNA gene sequences of strain LH_867 and those of 69 *Bifidobacterium* type strains were aligned using the SILVA Incremental Aligner (SINA aligner) v1.2.11 (*311*) and trimmed for quality using trimAl v1.4.1 (*312*) with -automated1 option optimised for maximum likelihood phylogenetic analyses, resulting in multiple sequence alignment consisting of 1496 positions. Models of nucleotide sequence evolution were evaluated using jModelTest2 v2.1.0 (*313*) under corrected Akaike Information Criterion (AICc) (*314*) and based on the results an appropriate model (GTR+G) was selected for phylogeny estimation. FastTree v2.1.9 with 1000 bootstrap iterations was used for phylogeny reconstruction (*315*). *Scardovia inopinata* JCM 12537^{T} (=DSM 10107^{T}) was used as an outgroup.

Additionally, sequences for *rpoB, rpoC, groL, dnaJ, clpC, dnaG* and *xpf* genes were *in silico* extracted from all annotated genomes and concatenated for the purpose of multilocus sequence analysis. MAFFT v7.450 (*316*) was used to create the multiple sequence alignment, which was then subjected to quality trimming and model selection as above. GTR+G model was selected for the phylogenetic tree

reconstruction using FastTree v2.1.9 with 1000 bootstrap iterations. The tree was rooted using *Scardovia inopinata* JCM 12537^{T} (=DSM 10107^{T}).

Furthermore, an alignment-free composition vector approach was used to reconstruct whole genome-based phylogeny from protein sequences employing a standalone version of CVTree v5.0 (*317*) with *k*-mer length of *k*=6.

2.4.7 Pangenomics, phylogenomics and comparative analyses

2.4.7.1 Breast- and formula-fed infant study (*B. longum*) (Chapter 3)

General feature format files of *B. longum* strains were inputted into the Roary pan-genome pipeline v.3.12.0 to obtain core-genome data and the multiple sequence alignment (msa) of core genes (MAFFT v7.313) (*318, 319*). All SNP analyses of strains from individual infants were performed using Snippy v4.2.1 (*320*) and the resulting msa was passed to the recombination removal tool Gubbins (*321*). Alignments resulting from all previous steps were cleaned from poorly aligned positions using manual curation and Gblocks v0.9b where appropriate (*322*). The core-genome tree was generated using FastTree v2.1.9 using the GTR model with 1000 bootstrap iterations. Snippy v4.2.1 with the "—ctgs" option, SNP-sites v2.3.3 (*323*) and FastTree v2.1.9 (GTR model with 1000 bootstrap iterations) were used to generate the whole genome SNP tree. Snp-dists v0.2 was used to generate pairwise SNP distance matrix between strains within individual infants (*324*). Scoary v1.6.16 with Benjamini Hochberg correction (*325*) was used on the output file from Roary to associate subsets of genes with specific traits – breast-fed, formula-fed, pre-weaning, weaning and post-weaning. The p-value cut-off was set to <1e-5, sensitivity cut-off to ≥70 % and specificity cut-off to ≥90 % to report the most overrepresented genes.

2.4.7.2 Wild mammal study (Chapter 4)

Anvi'o version 6.1 (*326*) was used to generate pan-genomes and single copy core gene data for other analyses. Briefly, a text file was created containing required

information on the collection of genomes and this file was used to generate genomes storage database. Next, the pan-genome was computed using the script anvi-pan-genome with parameters "--minbit 0.5 --mcl-inflation 10 --use-ncbi-blast". Single copy core genes were identified in the pan-genome and their aligned amino acid sequences recovered using anvi-get-sequences-for-gene-clusters with parameters "--min-num-genomes-gene-cluster-occurs", "--max-num-genes-from-each-genome", "--concatenate-gene-clusters". The resulting output was cleaned from poorly aligned positions using trimAl v1.4.1 (gaps in more than 50% of the genes) (*312*). IQ-TREE v1.6.1 (*327*) employing the 'WAG' general matrix model with 1000 bootstrap iterations (*328*) was used to infer the maximum likelihood trees from protein sequences. Script anvi-import-misc-data was used to import the results into the anvi'o pan-genome database, and the output was visualised with anvi-display-pan.

The host tree for co-phylogentic analysis for the wild mammal study was constructed using concatenated nucleotide sequences for 12S rRNA and partial cytb genes aligned using MAFFT v7.450, employing the 'GTR' model with 1000 bootstrap iterations. ParaFit (*329*) in the 'ape' package of R (*329, 330*) was used for topology-based comparisons.

2.4.7.3 Captive animal study (Chapter 5)

Sequences for *rpoB, rpoC, groL, dnaJ, clpC* and *xpf g*enes were *in silico* extracted from annotated genomes of isolates belonging to potentially novel *Bifidobacterium* species, as well as the 87 recognised *Bifidobacterium* type strains, and concatenated for the purpose of multilocus sequence analysis. MAFFT v7.450 (*316*) was used to create the multiple sequence alignment, which was then subjected to quality trimming with trimAl v1.4.1 (gaps in more than 50% of the genes) (*312*) and tree generation using IQ-TREE v1.6.1 (*327*) employing the 'GTR' model with 1000 bootstrap iterations (*328*).

General feature format files of *Bifidobacterium* strains were inputted into the Roary pan-genome pipeline v.3.12.0 with the BLASTP threshold set to 50% to obtain core-

genome data and the multiple sequence alignment (msa) of core genes (MAFFT v7.313) (*318, 319*). The resulting msa was processed as above, to generate the maximum likelihood core-genome tree.

2.4.8 CAZyme analysis

Functional categories (COG categories) were assigned to genes using EggNOG-mapper v0.99.3, based on the EggNOG database (bacteria) (*331*) and the abundance of genes involved in carbohydrate metabolism was calculated. A standalone version of dbCAN2 (v2.0.1) was used for glycobiome annotation (*332*). Glycosyl hydrolase (GH) gain-loss events were predicted using Dollo parsimony implemented in Count v9.1106 (*333*).

2.4.9 Screening for the presence of *eps* genes

BLAST+ v2.9.0 (e-value of 1e-5 and 50% identity over 50% sequence coverage) (*334*) was used to screen *B. castoris* genomes for the presence of homologues of *eps* genes from *B. animalis* subsp. *lactis* Bl12 (accession number CP004053.1, *eps*3: Bl12_1287 – Bl12_1328) and *B. pseudolongum* subsp. *globosum* LMG 11569^{T} (accession number JGZG01000015.1, *eps*4: BPSG_1548 – BPSG_1565).

2.4.10 Horizontal gene transfer prediction

SIGI-HMM (*335*) tool implemented in Islandviewer4 (*336*) was employed to predict HGT events.

2.4.11 CRISPR-Cas prediction

CRPSPRCasFinder v1.1.2 (*337*) was used to predict the presence of CRISPR-Cas loci. Sequences for CRISPR arrays were retrieved using CRISPRDetect v2.4 (*338*) and visualised using CRISPRStudio v1.0 (*339*).

2.4.12 Prophage prediction

Prediction of prophage sequences was made using VirSorter v1.0.6 (*340*). The software identifies viral regions in bacterial genomes and categorises them as follows: 1 and 4 (viral and lysogen, respectively) are “most confident” predictions similar to known viral references, categories 2 and 5 are “likely” predictions divergent from references, as well as partial sequences lacking viral hallmark genes which may include defective prophages, while categories 3 and 6 comprise sequences or regions with structure similar to viral genomes, but lacking similarity to known viruses or viromes. Sequences predicted to belong to categories 3 and 6 were excluded from analyses. Prophage sequences were visualised using BRIG v0.95 (*341*). Whole predicted prophage sequences, as well as the sequence for phage signature portal protein were used to construct maximum likelihood phylogenetic trees in IQ-TREE v1.6.1 employing ‘GTR’ and ‘WAG’ models, respectively. BLAST+ v2.9.0 with options “-evalue 1 -gapopen 10 -gapextend 1” was used to screen spacers identified in strains with complete CRISPR-Cas systems against predicted prophage sequences and *Bifidobacterium* genomes. Only matches showing 100% identity and 100% coverage were retained.

2.4.13 Nucleotide sequence accessions

Genome assemblies for the breast- and formula-fed infant study (*B. longum*) (Chapter 3) have been deposited at the GOLD database (https://img.jgi.doe.gov) under the accession number Gs0145337.

2.5 Experimental methods

2.5.1 Carbohydrate utilisation assay

Bifidobacterium (1%, v/v) was grown in modified (m)MRS (pH 6.8) supplemented with cysteine HCl at 0.05% and 2% (w/v) of selected carbohydrates as described

previously (*162*), except for pectin and mucin which were added at 1% (w/v). Growth was determined over a 48-h period using Tecan Infinite 50 microplate spectrophotometer at OD_{595}. Experiments were performed in biologically independent triplicates, and the plate reader measurements were taken automatically every 15min following 60s of shaking at normal speed. Due to the drop in initial OD values (i.e. recorded between T_0 and T_1) growth data were expressed as mean of the replicates between T_2 (30 min) and T_{end} (48-h).

	Item	Supplier
mMRS	Tripticase peptone	BD Biosciences, New Jersey, USA
	Yeast extract	Oxoid (Thermo Fisher Scientific), Loughborough, UK
	Tryptose	BD Biosciences, New Jersey, USA
	K2HPO4	Sigma-Aldrich, Dorset, UK
	KH2PO4	Sigma-Aldrich, Dorset, UK
	Ammonium citrate tribasic	Sigma-Aldrich, Dorset, UK
	Pyruvic acid	Sigma-Aldrich, Dorset, UK
	Cysteine HCl	Sigma-Aldrich, Dorset, UK
	Tween 80	Sigma-Aldrich, Dorset, UK
	MgSO4 • 7H2O	Sigma-Aldrich, Dorset, UK
	MnSO4 • 4H2O	Sigma-Aldrich, Dorset, UK
	FeSO4 • 7H2O	Sigma-Aldrich, Dorset, UK
Carbon sources	Lactose monohydrate	Sigma-Aldrich, Dorset, UK
	D-(+)-Glucose	Sigma-Aldrich, Dorset, UK
	D-(+)-Cellobiose	Sigma-Aldrich, Dorset, UK
	D-(+)-Mannose	Sigma-Aldrich, Dorset, UK
	Mucin from porcine stomach	Sigma-Aldrich, Dorset, UK
	Pectin from apple	Sigma-Aldrich, Dorset, UK
	L-Rhamnose monohydrate	Sigma-Aldrich, Dorset, UK
	D-(+)-Xylose	Sigma-Aldrich, Dorset, UK
	L-(+)-Arabinose	Sigma-Aldrich, Dorset, UK
	(+)-Arabinogalactan	Sigma-Aldrich, Dorset, UK
	Lacto-N-neotetraose	Glycom, Hørsholm, Denmark
	2'-fucosyllactose	Glycom, Hørsholm, Denmark

Table 2.8 Materials used in carbohydrate utilisation assays.

2.5.2 High-performance anion-exchange chromatography (HPAEC)

Mono-, di- and oligo- saccharides present in the spent media samples were analyzed on a Dionex ICS-5000 HPAEC system operated by the Chromeleon software version 7 (Dionex, Thermo Scientific). Samples were bound to a Dionex CarboPac PA1 (Thermo Scientific) analytical column (2 × 250 mm) in combination with a CarboPac PA1 guard column (2 × 50 mm), equilibrated with 0.1 M sodium hydroxide (NaOH). Carbohydrates were detected by pulsed amperometric detection (PAD). The system was run at a flow rate of 0.25 mL/min. The separation was done using a stepwise gradient going from 0.1 M NaOH to 0.1 M NaOH–0.1 M sodium acetate (NaOAc) over 10 min, 0.1 M NaOH–0.3 M NaOAc over 25 min

followed by a 5 min exponential gradient to 1 M NaOAc, before reconditioning with 0.1 M NaOH for 10 min. Commercial glucose, cellobiose, fucose, lactose and lacto-*N*-neotetraose (LNnT) were used as external standards. This experiment was performed in collaboration with Dr Sabina Leanti La Rosa, who also analysed the HPAEC data.

2.5.3 Proteomics

B. longum subsp. *longum* strain 25 (B_25) was grown in triplicate in mMRS supplemented with L-cysteine HCl at 0.05% and 2% (w/v) glucose, cellobiose or LNnT as a sole carbon source. *B. longum* subsp. *longum* strain 71 (B_71) was grown in triplicate in mMRS supplemented with cysteine HCl at 0.05% and either 2% (w/v) glucose or 2'-FL as a sole carbon source. Cell pellets from 50 mL samples (at the mid-exponential growth phase) were collected by centrifugation (4500 × g, 10 min, 4 °C) and washed three times with PBS pH7.4. Cells were resuspended in 50mM Tris-HCl pH8.4 and disrupted by bead-beating in three 60s cycles using a FastPrep24 (MP Biomedicals, CA). Protein concentration was determined using a Bradford protein assay (Bio-Rad, Germany). Protein samples, containing 50µg total protein, were separated by SDS-PAGE with a 10% Mini-PROTEAN gel (Bio-Rad Laboratories, CA) and then stained with Coomassie brilliant blue R250. The gel was cut into five slices, after which proteins were reduced, alkylated, and in-gel digested as previously described (*342*). Peptides were dissolved in 2% acetonitrile containing 0.1% trifluoroacetic acid and desalted using C18 ZipTips (Merck Millipore, Germany). Each sample was independently analysed on a Q-Exactive hybrid quadrupole-orbitrap mass spectrometer (Thermo Scientific) equipped with a nano-electrospray ion source. MS and MS/MS data were acquired using Xcalibur (v2.2 SP1). Spectra were analysed using MaxQuant v1.6.1.0 (*343*) and searched against a sample-specific database generated from the B_25 and B_71 genomes. Proteins were quantified using the MaxLFQ algorithm (*344*). The enzyme specificity was set to consider tryptic peptides and two missed cleavages were allowed. Oxidation of methionine, N-terminal acetylation and deamidation of asparagine and glutamine and formation of pyro-glutamic acid at N-terminal glutamines were used as variable

modifications, whereas carbamidomethylation of cysteine residues was used as a fixed modification. All identifications were filtered in order to achieve a protein false discovery rate (FDR) of 1% using the target-decoy strategy. A protein was considered confidently identified if it was detected in at least two of the three biological replicates in at least one glycan substrate. The MaxQuant output was further explored in Perseus v1.6.1.1 (*345*). The proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with dataset identifier PXD017277. This experiment was performed in collaboration with Dr Sabina Leanti La Rosa, who also carried out the analysis of the proteomics data.

2.6 Graphs and illustrations

Phylogenetic trees were visualised using iTOL (*346*). RStudio v1.3.1093 was used for R programming (*347*). Figures were generated using R packages 'gplots' 'ComplexHeatmap', 'gggenes' and GraphPad Prism v8.4.3 for MacOS, GraphPad Software, La Jolla California USA (www.graphpad.com). Inkscape v1.0.0beta for MacOS was used to edit figures.

2.7 Statistical analyses

For all statistical analyses, *P* values <0.05 were considered to be significant.

2.7.1 Breast-fed and formula-fed infant study (*B. longum*) (Chapter 3)

The T-test function implemented in Microsoft Excel v16.16.20 was used to calculate statistically significant differences between average numbers of GH genes belonging to the predominant GH families.

2.7.2 Wild mammal study (Chapter 4)

χ^2 function implemented in R was used to assess difference in prevalence of *Bifidobacterium* between wild mice and voles.

2.7.3 Captive animal study (Chapter 5)

Normality of data was assessed using Shapiro-Wilk normality test implemented in GraphPad Prism v8.4.3. To account for the non-normal data and non-equal sample sizes, the Kruskal-Wallis and the Dunn's post hoc tests (GraphPad Prism) were used to compare genome size and GH abundance between *Bifidobacterium* isolates belonging to different host categories. The T-test (GraphPad Prism) was used for comparisons between infant and adult categories within the human host group (normal data distribution). ANOSIM function implemented in R was used to test whether the host categories were associated with phylogenetic relatedness of *Bifidobacterium* based on the distance matrix generated from the maximum likelihood core-genome tree using 'cophenetic' function in 'ape' package.

Chapter 3

Succession of *Bifidobacterium longum* strains in response to a changing early life nutritional environment reveals dietary substrate adaptations.

Dr Anne McCartney performed DNA extraction and provided strain collection.

DNA library preparation for WGS was done by sequencing teams at the Wellcome Sanger Institute (Hinxton, UK) and Quadram Institute Bioscience (Norwich, UK).

I performed carbohydrate growth experiments and prepared selected samples for carbohydrate uptake assay (HPAEC) and proteomics.

Carbohydrate uptake and proteomics were performed by both Dr Sabina Leanti La Rosa and myself.

I performed all computational analyses, except for the analysis of the results from HPAEC and proteomics, which was carried out by Dr Sabina Leanti La Rosa.

This chapter has been published in:

Kujawska, M. *et al.*, Succession of *Bifidobacterium* longum Strains in Response to a Changing Early Life Nutritional Environment Reveals Dietary Substrate Adaptations. *iScience* 23, 101368 (2020).
DOI:https://doi.org/10.1016/j.isci.2020.10136

3.1 Introduction

Diet-microbe interactions at early life stages are crucial for infant development and microbiota modulation. The gut of breast-fed infants is primarily colonised by members of genus *Bifidobacterium*, with particularly high prevalence of strains belonging to *B. longum* subsp. *longum* (*B. longum*) and *B. longum* subsp. *infantis* (*B. infantis*). Although introduction of a more diversified diet later in infancy initiates a shift in microbiota from *Bifidobacterium*-dominated to a more complex one, specific strains of *B. longum* can persist long-term in individual hosts. This study aimed to investigate the adaptation of *B. longum* to the changing infant diet. Examination of the genomes of 75 strains isolated from nine infants in their first 18 months (either exclusively breast- or formula-fed pre-weaning), indicated intra-individual and subspecies-specific strain diversity with respect to glycosyl hydrolase families and enzymes, which corresponded to different dietary stages. Phenotypic growth studies revealed strain-specific differences in utilisation of human milk oligosaccharides and plant carbohydrates between and within individual infants. Furthermore, carbohydrate uptake assay and proteomic profiling identified active gene clusters involved in degradation of selected carbohydrates. The results indicate a strong link between infant diet and the genomic diversity of *B. longum* subspecies and strains reflected in their carbohydrate utilisation profiles, which is in line with changes to the nutritional environment: i.e. moving from breast milk to a more complex diet. These data provide additional insights into possible mechanisms behind the competitive advantage of *B. longum* and its long-term persistence in a single host and may contribute to rational development of new dietary therapies for this important developmental window.

3.2 Background

The mutualistic relationship between the host and its microbiota begins with microbial colonisation shortly after birth (*70, 162, 348*). Infant development is strongly linked to the early life microbiota, with its crucial role in modulating immune responses, providing resistance to pathogens, and digesting dietary components (*59, 132, 153, 349-352*). Indeed, interactions between host diet and gut microbes are suggested to play a central role during infancy, leading to health effects that may extend to later life stages (*77-80, 83, 85*). The microbiota of vaginally delivered full-term healthy infants is relatively simple and characterised by the dominance of the genus *Bifidobacterium* (*353, 354*). In contrast, disrupted transmission of maternal gastrointestinal bacteria, such as *Bifidobacterium*, in infants born through caesarean section results in high levels of opportunistic hospital-associated pathogens, which have been shown to account for more than 65% of the total microbiota composition in the first 4 days after birth (*354*).

Considered the gold nutritional standard for infants, breast milk is also an important dietary supplement for early life microbiota members, including *Bifidobacterium*. The strong association between infant diet and gut microbes has further been supported by reports of differences in microbiota composition between breast- and formula-fed infants, linking low abundance of *Bifidobacterium* to formula-feeding (*92, 172*). These microbiota differences have further been associated with differential health outcomes between the two groups, for example increased instances of asthma, allergy and obesity in formula-fed infants (*174, 178-180, 355*).

The presence of specific carbohydrate utilisation genes and gene clusters in genomes of *Bifidobacterium*, particularly the ones involved in the metabolism of breast milk-associated human milk oligosaccharides (HMOs), has been linked to high abundance of *Bifidobacterium* in breast-fed infants (*59*). These genes often display species- and strain-specificity, and their presence has been described in *B. breve*, *B. bifidum*, *B. longum*, *B. infantis*, and more rarely in *B. pseudocatenulatum (59, 156-158).* Additionally, co-existence of *Bifidobacterium*

species and strains in individual infant hosts, and resulting interactions and metabolic co-operation within a single (HMO-associated) ecosystem have been reported (*38, 162*).

Transition from breast milk to a more complex diet, coupled with the introduction of solid foods, has been considered the first step in the development of an adult-like microbiome, functionally more complex and harbouring genes responsible for degradation of plant-derived complex carbohydrates, starches, and xenobiotics, as well as production of vitamins (*88, 241*). Non-digestible complex carbohydrates, in particular, including arabinoxylans (AX), inulin-type fructans (ITF) or arabinoxylo-oligosaccharides (AXOS) present in complementary foods have been suggested to potentially exert beneficial health effects on hosts via their bifidogenic and prebiotic properties and resulting modulation of the intestinal microbiota and metabolic end-products (*55, 242-244*).

Specific strains of *Bifidobacterium*, and *B. longum* in particular, have previously been shown to persist in individual infant hosts long-term, despite changes in microbiota composition during weaning (*356, 357*). Currently, *B. longum* is recognised as four subspecies, with subspecies *longum* and *infantis* associated with the human gut microbiota, and subspecies *suis* and *suillum* isolated from animal hosts (*358, 359*). It is considered to be the most common and prevalent species in the human gut, with *B. longum* subsp. *infantis* associated with infants, and *B. longum* subsp. *longum* linked to both infants and adults (*24, 72*). The differences in prevalence between the two subspecies, and the ability of human host to acquire new *B. longum* strains at different life stages have been associated with particular bacterial carbohydrate utilisation capabilities and the overall composition of the resident microbiota (*289, 360*).

Recently, several metagenomic studies have investigated the early life microbiota in breast- and formula-fed infants (*296, 361-363*). The investigation of the DIABIMMUNE cohort, employing strain-level metagenomics, provided insights into diet-related functional aspects of *B. infantis* in breast-fed infants (*288*). Although several longitudinal studies focused specifically on *B. longum* and reported intraspecies diversity, colonisation and long-term persistence (years) of this species

in hosts (*289, 357, 364*), investigations into diet-related functions at early life stages remain limited. The studies examining infant-associated *B. longum* strains in relation to diet have not been profiled over longitudinal and changing dietary periods (*365*). Hence, further detailed longitudinal studies are required to assess *B. longum* strains in single hosts over time, with a particular focus on changing dietary patterns.

In this study, I investigated the adaptations of *Bifidobacterium* to the changing infant diet and examined a unique collection of *B. longum* strains isolated from nine either exclusively breast- or formula-fed infants across their first 18 months, encompassing pre-weaning, weaning and post-weaning dietary stages. I assessed the genomic and phenotypic carbohydrate metabolism similarities between 62 *B. longum* strains and 13 *B. infantis* strains. The results indicate a strong association between host diet and *Bifidobacterium* species and strains, which seems to align with changes to the nutritional environment.

3.3 Hypothesis and aims

Whole genome sequences of *Bifidobacterium longum* display particular genomic features related to carbohydrate metabolism. These features may play a role in the establishment and persistence of members of this species in individual hosts during infancy despite dietary changes associated with transition from breast milk to solid foods.

Aims:

1) Explore genomic features of *B. longum* strains isolated from breast- and formula-fed infants during the first 18 months of life, with particular focus on traits related to carbohydrate metabolism
2) Perform phenotypic growth experiments to assess functional capabilities of representative *B. longum* isolates. Link genomic predictions to phenotypic data
3) Perform further experimental characterisation of strains displaying interesting phenotypic properties

3.4 Results

Previously, studies examining *B. longum* across the human lifespan have reported a broad distribution of this species and have suggested their ability to persist within hosts for prolonged periods of time and through changing nutritional environment (*356, 357*). To better understand potential mechanisms facilitating these properties during the early life window, I sought to investigate the genotypic and phenotypic traits of *B. longum* strains within individual infants in relation to diet (i.e. breast milk vs formula) and dietary stages (i.e. pre-weaning, weaning and post-weaning), following up on a longitudinal study of the infant faecal microbiota published in 2010 (*296*). Briefly, faecal samples from seven either exclusively breast- or formula-fed infants were collected regularly from 1 month to 18 months of age (*296*). The number of samples obtained from the breast-fed infants during the pre-weaning period was higher than that obtained from the formula-fed group, which may correlate with the age at which the two groups started weaning (~20.6 vs. ~17 weeks old).

In the original study, quantitative analysis using fluorescence in situ hybridization (FISH) was performed on collected samples to enumerate the predominant bacterial groups (Table S3.1) (*296*). In addition, bacterial isolation was carried out on selected samples, and the isolated colonies identified using ribosomal intergenic spacer analysis (*296*).

3.4.1 Quantitative analysis of microbial communities in breast- and formula-fed infants

I first sought to provide context to the gastrointestinal environment *B. longum* strains selected for the present study were isolated from. For this purpose, I re-analysed the data originally generated by FISH (Figure 3.1 & Table S3.1) (*296*)). Proportionally, bifidobacteria detected with probe Bif164 constituted the predominant group in samples from breast-fed infants during pre-weaning and weaning, with their levels fluctuating from 16.5% to 100% of the microbiota across these dietary stages. Proportions of bifidobacteria across all breast-fed samples

decreased considerably during post-weaning and ranged from 4.6% to 12.1%. The levels of bacteria detected by Erec482 (members of *Clostridium* cluster XIVa) started to increase during weaning in all samples and this trend continued until the end of the study (mean±sd 0.25±0.66%, 2.04±3.39% and 18.16±10.24 % for pre-weaning, weaning and post-weaning phases, respectively). Members of genus *Bacteroides*, *Parabacteroides* and *Prevotella* species, *Paraprevotella*, *Xylanibacter*, *Barnesiella* species and *Odoribacter splanchnicus*, detectable by probe Bac303, were identified in all samples throughout the study. This bacterial group showed extensive inter-individual variation, with mean proportions of 3.29±6.53%, 4.9±10.35% and 13.1±9.73% for pre-weaning, weaning and post-weaning phases. Other microbiota members were detected in breast-fed samples at lower levels, including members of family *Coriobacteriia* (Ato291, mean <2 % of microbiota), *Escherichia coli* (EC1531, <1 %), members of *Clostridium* clusters I and II (Chis150, <1 %) and lactic acid bacteria (Lab158, mean <1 %).

Figure 3.1 Proportional representation of bacterial populations in the faecal microbiota of a) breast-fed and b) formula-fed infants based on FISH analysis.
Numbers are expressed as percentage of the total bacterial population obtained using DAPI. The vertical solid black lines mark the different dietary phases in each infant (pre-weaning, weaning and post-weaning). Oligonucleotide probes used to determine bacterial populations: ***Bif164*** *– most Bifidobacterium species and Parascardovia denticolens,* ***Bac303*** *– most members of the genus Bacteroides, some Parabacteroides and Prevotella species, Paraprevotella, Xylanibacter, Barnesiella species and Odoribacter splanchnicus,* ***Erec482*** *– most members of Clostridium cluster XIVa,* ***Ato291*** *– Cryptobacterium curtum, Gordonibacter pamelaeae, Paraeggerthella hongkongensis, all Eggerthella, Collinsella, Olsenella and Atopobium species,* ***Chis150*** *– most members of Clostridium cluster I, all members of Clostridium cluster II,* ***EC1531*** *– Escherichia coli,* ***Lab158*** *– all Oenococcus, Vagococcus, Melissococcus, Tetragenococcus, Enterococcus, Catellicoccus, Paralactobacillus, Pediococcus and Lactococcus species, most Lactobacillus, Weissella and Leuconostoc species.*

In contrast to the breast-fed group, no drastic changes in bacterial populations were observed in formula-fed infants throughout the study. Overall, levels of bifidobacteria detected during pre-weaning and weaning in this group were lower and fluctuated from 0.0% to 73.3% of the microbiota at different time points. During post-weaning, proportions of *Bifidobacterium* decreased across all formula-fed samples and ranged from 6.5% to 12% at month 18, similar to the breast-fed group. The levels of bacteria detected by probe Erec482 were overall higher in formula-fed samples throughout the duration of the study, with mean proportions of 19.96±17.41%, 25.39±14.63% and 30.6±15.92% for pre-weaning, weaning and post-weaning phases. Similarly, proportions of bacteria belonging to bacteroides group (probe Bac303) were higher in the formula-fed group compared to the breast-fed group during all dietary phases, with mean proportions of 3.55±5.51%, 13.25±13.78% and 25.73±19.04% for pre-weaning, weaning and post-weaning, respectively. In contrast to the breast-fed group, levels of bacteria belonging to *Clostridium* clusters I and II (probe Chis150) started to increase during weaning in the formula-fed group and continued to increase throughout the remainder of the study (1.23±1.28%, 7.03±9.18% and 21.72±11.47% for pre-weaning, weaning and post-weaning, respectively). Levels of bacteria identified by Ato291 and EC1531 in formula-fed samples were slightly higher than in the breast-fed group (means of <3.5% and <1.25%, respectively), while the mean proportion of lactic acid bacteria (Lab158) remained below <1%.

Overall, these results are in line with previous reports of differences in faecal microbiota composition between breast- and formula-fed babies, in particular during the pre-weaning and weaning phases, and demonstrate the succession of bacterial species over time and in relation to diet, including *Bifidobacterium*.

3.4.2 General features of *B. longum* genomes

Based on the previously published results of bacterial culture and colony identification (for details, refer to (*363*)), I selected 88 isolates that were originally identified as *Bifidobacterium* for the present study. Forty-six strains were recovered

from five exclusively breast-fed infants (BF1-BF5, including identical twins BF3 and BF4) and further 42 from four exclusively formula-fed infants (FF1-FF3 and FF5). Following sequencing and based on the initial ANI analysis (Tables S3.2 & S3.3), 75 strains were identified as *B. longum* spp. and included in further analyses, with 62 strains identified as *B. longum* subsp. *longum* (*B. longum*) and 13 strains identified as *B. longum* subsp. *infantis* (*B. infantis*) (Figure 3.2).

To determine potential genotypic factors facilitating establishment and persistence of *B. longum* in the changing early life environment, I sought to assess the genome diversity of the selected strains. Sequencing generated between 12 and 193 contigs for each *B. longum* strain, with 74 out of 75 genomes containing fewer than 70 contigs, yielding a mean of 66.95-fold coverage (Table S3.2). The predicted genome size for strains identified as *B. longum* ranged from 2.21 Mb to 2.58 Mb, with an average G+C% content of 60.11%, an average predicted ORF number of 2,023 and number of tRNA genes ranging from 55-88. For strains identified as *B. infantis*, the predicted genome size ranged from 2.51 Mb to 2.75 Mb, with an average G+C% content of 59.69%, an average predicted ORF number of 2,280 and the number of tRNA genes ranging from 57 to 62.

3.4.3 Comparative genomics

To assess nucleotide-level genomic differences between *B. longum* strains, I performed ANI analysis. Results (Table S3.3) indicated higher levels of sequence identity between *B. longum* strains isolated from individual infant hosts than between those isolated from different hosts. For example, pairwise identity values for strains isolated from infant BF3 showed the narrowest range (average value of 99.99±3.15e-5%), followed by infant FF2 strains (99.98±1.12e-4%), with infant BF2 strains having the broadest identity value range (averaging 99.13±7.8e-3%).

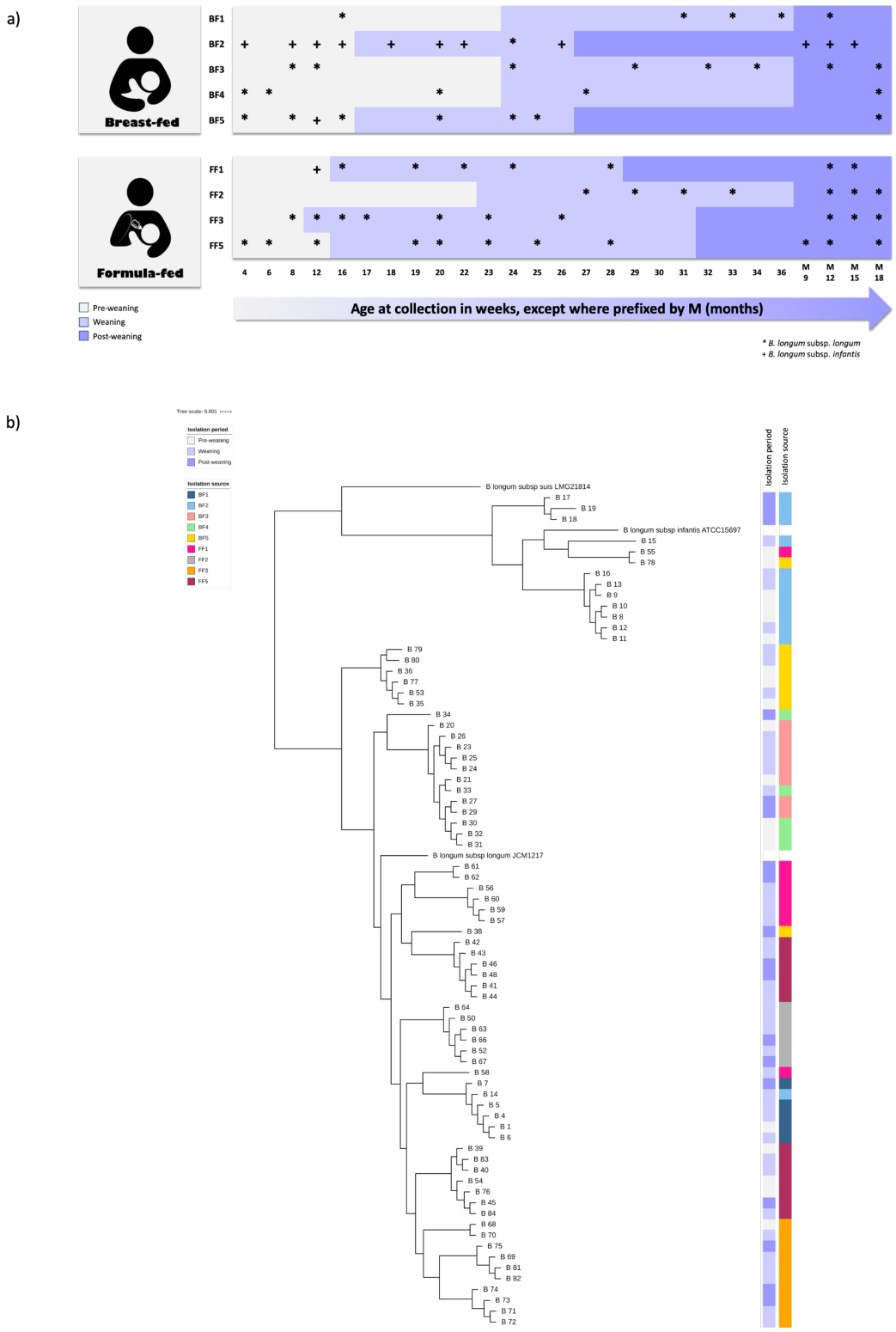


Figure 3.2 Identification and relatedness of B. longum strains.
Sampling scheme and strain identification within individual breast-fed (BF1-BF5) and formula-fed (FF1-FF3 and FF5) infants based on average nucleotide identity values (ANI). The three levels of shading mark different dietary phases: pre-weaning, weaning, and post-weaning. B) Relatedness of B. longum strains based on core proteins. Coloured strips represent isolation period (pre-weaning, weaning and post-weaning) and isolation source (individual infants), respectively.

Next, I examined genetic diversity of newly sequenced *B. longum* strains and their relatedness to each other, alongside the *B. longum* type strains *B. longum* subsp. *longum* JCM 1217^T, *B. longum* subsp. *infantis* ATCC 15697^T and *B. longum* subsp. suis LMG 21814^T. This analysis identified a total of 1002 core genes present in at least 99% of the analysed *B. longum* subspecies genomes. Based on the absence/presence of specific genes, a clear distinction between *B. longum* subspecies (i.e. *longum* vs. *infantis*) could be made. (Table S3.4). Phylogenetic analysis revealed the clustering of the *B. longum* strains within each subspecies according to isolation source (i.e. individual infants), rather than dietary stage (i.e. pre-weaning, weaning and post-weaning) (Figure 3.2). Interestingly, strains isolated from formula-fed baby FF5 clustered into two separate clusters, irrespective of the isolation period, suggesting that two highly related *B. longum* groups are present within this infant. Moreover, strains isolated from identical twins BF3 and BF4 clustered together, indicating their close relatedness.

Based on the generated pan-genome data, I next sought to identify whether specific components of the *B. longum* subspecies genomes were enriched in infant hosts using Scoary. This software sequentially scores each candidate gene in the accessory genome according to its apparent correlation to predefined traits, in this case host diet (breast vs. formula) or dietary stage (pre-weaning, weaning and post-weaning). Based on the results of this analysis, a gene annotated as α-L-arabinofuranosidase, along with four other genes coding for hypothetical proteins, were predicted to be overrepresented in *B. longum* strains isolated from breast-fed infants. Alpha-L-arabinofuranosidases catalyse hydrolysis of terminal non-reducing α-L-arabinofuranoside residues in α-L-arabinosides and act on such carbohydrates as (arabino)xylans (*366, 367*). In addition, two genes encoding hypothetical proteins and a gene coding for Mobility protein A were overrepresented in strains isolated from formula-fed infants. Furthermore, no associations between genes and dietary stages in *B. longum*, nor any associations whatsoever in *B. infantis* were observed (Table S3.5).

Given that *B. longum* strains were isolated from individual infants at different time points, I next sought to assess their intra-strain diversity. I used the first *B. longum*

isolate from each infant as the “reference” strain, to which all other strains from the same infant were compared (Figure 3.3). The lowest strain diversity was observed in infants BF1, BF3 and FF2, with respective mean pairwise SNP distances of 18.7±20.3 SNPs (mean±sd), 10.3±5.0 SNPs and 13.3±5.3 SNPs. These results indicate that strains isolated from these infants may be clonal, suggesting long-term persistence despite dietary changes. Surprisingly, analysis of strains isolated from breast-fed identical twins BF3 and BF4 revealed higher strain diversity in baby BF4 (1034.5±1327.1 SNPs), compared to the highly similar strains in infant BF3 (i.e. 10.3±5.0 SNPs).

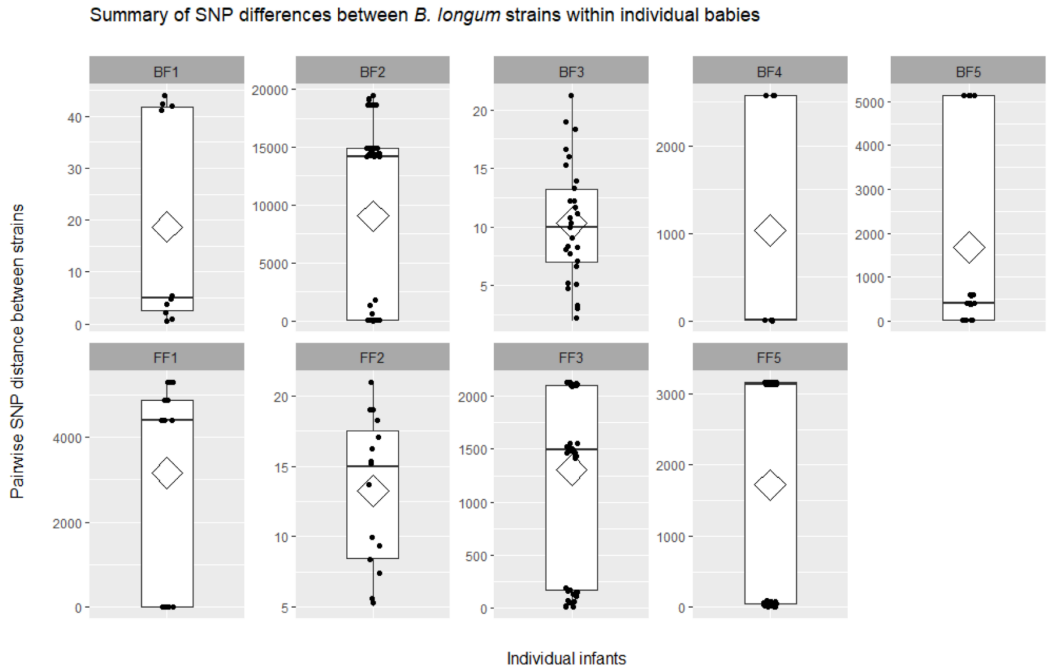


Figure 3.3 Pairwise SNP distances between B. longum strains of the same subspecies within individual infants.
Individual points show data distribution, diamonds indicate the group mean, box plots show group median and interquartile range.

Based on these results, I conducted SNP analysis on *B. longum* strains isolated from both babies. The data indicated that out of 13 strains analysed (n=8 from BF3 and n=5 from BF4), 12 isolated during pre-weaning, weaning and post-weaning seemed to be clonal (with mean pairwise SNP distance of 10.0±5.5 SNPs) and one strain from baby BF4 isolated post-weaning was more distant, 2595.4±2.8 SNPs. This difference in strain diversity may possibly be explained by the fact that infant BF4 received a

course of antibiotics during pre-weaning (Figure 3.3, Tables S3.2 and S3.6) (*296*). Furthermore, the presence of clonal strains in both babies suggests vertical transmission of maternal *B. longum* strains to both infants, or potential horizontal strain transmission between babies, consistent with previous reports (*146, 147, 289, 368*). *B. infantis* strains isolated from infant BF2 showed the highest strain diversity, with the mean pairwise distance of 9030.9±8036.6 SNPs. Seven strains isolated during both pre-weaning and weaning periods appeared to be clonal, 6.3±1.6 SNPs, while four strains isolated during weaning and post-weaning were more distant, with mean pairwise SNP distance of 14983.5±4658.3 SNPs (Table S3.6).

3.4.4 Functional annotation of *B. longum* subspecies genomes – carbohydrate utilisation

To investigate genomic differences between *B. longum* strains at a functional level, I next assigned functional categories to ORFs of each genome. This resulted in the identification of carbohydrate transport and metabolism as the second most abundant category (after unknown function), reflecting the saccharolytic lifestyle of *Bifidobacterium* (Figure S3.1) (*38, 42*). *B. longum* had a slightly higher proportion of carbohydrate metabolism and transport genes (11.39±0.31%) compared to *B. infantis* (10.20±0.60%), which is in line with previous reports (*56, 369*). *B. longum* strains isolated during pre-weaning had a similar proportion of carbohydrate metabolism genes in comparison with the strains isolated post-weaning: 11.28±0.23% and 11.48±0.38%, respectively. Furthermore, comparison between *B. longum* strains isolated from breast- and formula-fed infants produced similar results, with respective values of 11.41±0.21% and 11.38±0.38%. In contrast, *B. infantis* strains isolated pre-weaning had a lower proportion of carbohydrate metabolism genes in their genomes compared to the ones isolated post-weaning: 9.90±0.24% and 11.20±0.01%, respectively (Table S3.7).

Glycosyl hydrolases (GH), which facilitate glycan metabolism in the gastrointestinal tract, are one of the major classes of carbohydrate-active enzymes. I thus sought to investigate and compare the repertoire of GHs in *B. longum* using dbCAN2. This

analysis resulted in the identification of a total of 36 different GH families in all *Bifidobacterium* strains. *B. longum* was predicted to contain 55 GH genes per genome on average (2.72 % of OFRs), while this number was lower for *B. infantis* strains, with an average of 37 GH genes per genome (1.62% of ORFs) (Figure 3.4 & Table S3.8).

The predominant GH family within the *B. longum* group was GH43, whose members include enzymes involved in metabolism of complex plant carbohydrates such as (arabino)xylans (*370*), followed by GH13 (starch), GH51 (hemicelluloses) and GH3 (plant glycans) (*33, 38*). Strains isolated during pre-weaning had a slightly lower mean number of GH genes compared to strains isolated post-weaning (54.46±2.81 vs. 56.85±2.77). In addition, strains isolated from breast-fed babies contained an average of 53.96±3.82 GH genes per genome, while this number was slightly higher for strains isolated from formula-fed infants; 56.47±2.96. Further analysis revealed that these differences appeared to be intra-host-specific and related to diet. For example, strains isolated from breast-fed twins BF3 and BF4 pre-weaning were predicted to harbour 11 GH43 genes per genome, while the pre-weaning strain from formula-fed baby FF3 had 13 GH genes per genome predicted to belong to this GH family. Similarly, strains isolated from babies BF3 and BF4 post-weaning had 11 predicted GH43 genes, while the three strains isolated from infant FF3 were predicted to contain 16, 16 and 18 GH43 genes per genome, respectively (Table S3.8).

Next, I sought to determine whether observed differences in GH gene numbers statistically correlated with breast- and formula-fed groups (Table S3.8). The results of the *t-test* suggested significant differences ($P < 0.05$) between mean numbers of GH genes belonging to the predominant GH families (GH43 – higher abundance in FF babies, GH13 – higher abundance in BF babies, and GH51 – higher abundance in FF babies), and several other GH families, including GH5 (cellulases), GH38 (mannosylglycerate hydrolases) and GH36 (α-galactosidases), all more abundant in BF babies. Further analysis, with the focus on dietary phases, suggested significant differences in mean numbers of GH genes between breast- and formula-fed groups

during pre-weaning (e.g. families GH43, GH13, GH5, GH38), but not in the post-weaning phase (Table S3.8).

Given that glycosyl hydrolases belonging to distinct GH families may have similar catalytic properties, I next grouped the GH genes for which the predicted enzyme class annotation was available and investigated their abundance (Table S3.9). The predominant enzyme classes in *B. longum* strains were non-reducing end α-L-arabinofuranosidases belonging to GH43 and GH51, followed by β-galactosidases (GH2 and GH42), oligo-1,6-glucosidases (GH13) and β-*N*-acetylhexosaminidases (GH3 and GH20). As mentioned above, arabinofuranosidases (EC 3.2.1.55) hydrolyse α-1,2-, α-1,3-, α-1,5-L arabinofuranosidic bonds in arabinoxylan and L-arabinan and are accessory enzymes involved in the degradation of plant cell wall polysaccharides (*371*). The enzymatic activity of beta-galactosidases (EC 3.2.1.23) ranges from lactose present in breast milk to galacto-oligosaccharides and galactans found in plant cell walls (*372, 373*), while oligo-1,6-glucosidases (3.2.1.10) are debranching enzymes which act on oligosaccharides containing α-1,6 linkages to produce α-glucose (*374*). Beta-*N*-acetylhexosaminidases (EC 3.2.1.52) remove non-reducing terminal β-1,4 linked *N*-acetylglucosamine (GlcNAc) or β-*N*-acetylgalactosamine (GalNAc) residues of oligosaccharides and their conjugates (*375*).

The results of the analysis of mean numbers of enzyme classes between breast- and formula-fed babies suggested significant differences ($P < 0.05$) in the top three above-mentioned predominant enzyme classes as well as several other less abundant ones, including non-reducing end β-L-arabinofuranosidases (GH127 and GH146 – higher abundance in BF babies), α-galactosidases (GH36 – higher abundance in BF babies), and endo-1,5-α-L-arabinases (GH43 – higher abundance in FF babies). Additional analysis focusing on dietary phases suggested significant differences between breast- and formula fed groups during pre-weaning (e.g. non-reducing end α-L-arabinofuranosidases, β-galactosidases, oligo-1,6-glucosidases as well as α-galactosidases), but not during post-weaning (Table S3.9).

Next, I sought to examine the predicted glycosyl hydrolase repertoire of *B. infantis* strains, with the caveat that the majority of the strains belonging to this subspecies

were isolated from a single infant. The most abundant GH family in *B. infantis* was GH13 (starch), followed by GH42, GH20 and GH38 (Table S3.8), contrastingly to the *B. longum* group. *B. infantis* strains also harboured genes predicted to encode members of the GH33 family, which contains exo-sialidases (*38*). Strains isolated pre-weaning were predicted to contain an average of 34.83±0.4 GH genes per genome, while this number was higher for the strains isolated post-weaning (i.e. 43.00±0.00 GH genes). *B. infantis* strains isolated post-weaning contained families GH1 and GH43 that were absent in the strains isolated pre-weaning. The GH1 family contains enzymes such as β-glucosidases, β-galactosidases and β-D-fucosidases active on a wide variety of (phosphorylated) disaccharides, oligosaccharides, and sugar–aromatic conjugates (*376*).

The analysis of enzyme classes in the *B. infantis* strains suggested that β-galactosidases (GH2 and GH42) were predominant in this group, followed by β-*N*-actetylhexaminidases (GH3 and GH20), 4-α-glucanotransferases (GH77) and oligo-1,6-glucosidases (GH13) (Table S3.9). The enzyme 4-α-glucanotransferase (EC 2.4.1.25) has been reported to play a role in metabolism of maltose and maltodextrin by removing the reducing end glucosyl units from α-1,4-glucan and transfers the non-reducing dextrinyl moiety onto glucose or another α-1,4-glucan (*377*).

Genomes of members of the genus *Bifidobacterium* have previously been shown to contain GH genes involved in metabolism of various HMOs present in breast milk (*158, 378*). Genes belonging to GH29 and GH95 (α-L-fucosidases found active on fucosylated HMOs (*158, 379*)) were identified in all *B. infantis* strains, as well as four *B. longum* strains isolated from formula-fed baby FF3. Furthermore, all *B. infantis* and *B. longum* strains were predicted to harbour GH20 and GH112 genes (lacto-*N*-biosidases and galacto-*N*-biose/lacto-*N*-biose phosphorylases shown to be involved in degradation of isomeric lacto-*N*-tetraose (LNT) (*380*)) (Table S3.8).

Overall, these findings suggest differences in general carbohydrate utilisation between *B. longum* and *B. infantis* at different stages suggesting adaptation of *Bifidobacterium* to a changing early life diet, which may be a factor facilitating establishment of these bacteria within individual hosts during infancy.

3.4.5 Prediction of gain and loss of GH families in *B. longum*

Given the differences in the carbohydrate utilisation profiles between *B. longum* and *B. infantis*, I next sought to investigate the acquisition and loss of GH families using Count software. For this purpose, I additionally predicted the presence of GH families in type strains *B. longum* subsp. *longum* JCM 1217^T, *B. longum* subsp. *infantis* ATCC 15697^T and *B. longum* subsp. *suis* LMG 21814^T with dbCAN2 and generated a whole genome SNP tree to reflect gene loss/gain events more accurately (Figure 3.4 & Table S3.10). Both *B. longum* and *B. infantis* lineages appear to have acquired GH families (when compared to the common ancestor of the phylogenetic group), with the *B. longum* lineage gaining two GH families (GH121 and GH146) and the *B. infantis* lineage one GH family (GH33). Within the *B. infantis* lineage, which also contains the *B. suis* type strain, the *B. infantis* taxon has further acquired two and lost five GH families. These findings suggest that the two human-related subspecies have followed different evolutionary paths, which is in line with previous observation of differences between *B. longum* and *B. infantis* resulting from phylogenomic analyses (*365*). Interestingly, strain adaptation to the changing nutritional environment (i.e. individual infant gut) seems to be driven by loss of specific GH families (Figure 3.4). For example, *B. infantis* strains isolated during pre-weaning and weaning from baby BF2 appear to be missing up to three GH families (GH1, GH43 and GH109) present in strains isolated post-weaning. Lack of family GH43 associated with enzymes acting on plant-derived polysaccharides in early life *B. infantis* strains may explain nutritional preference of this subspecies for an HMO-rich diet.

Similarly, I observed differential gene loss events in *B. longum* strains from individual hosts. For example, all strains isolated from baby BF5 appear to lack GH families GH1, GH29 and GH95. However, strains isolated pre-weaning additionally lacked GH53 family, which includes endogalactanases shown to be involved in liberating

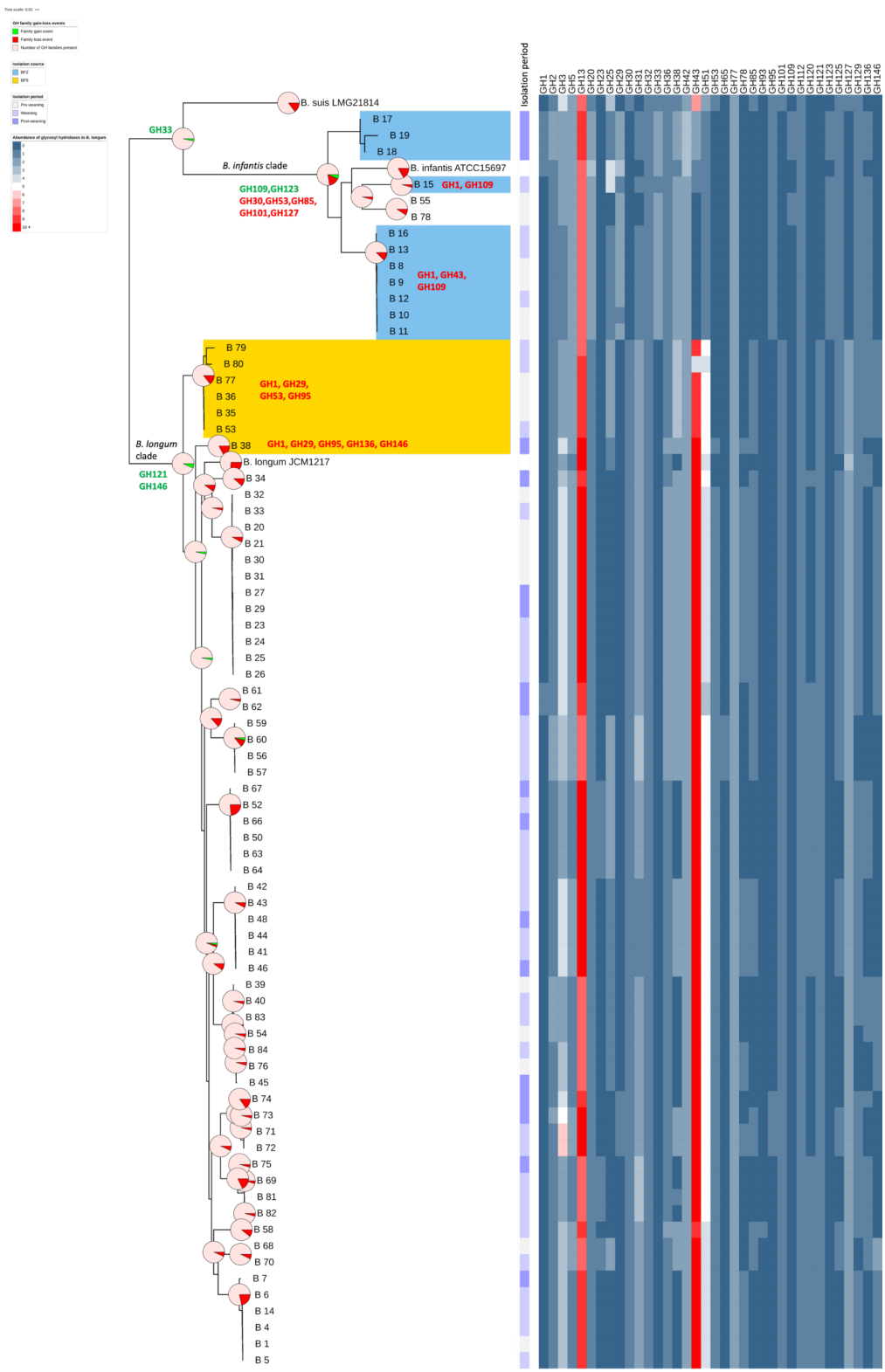


Figure 3.4 Gene-loss events and abundance of GH families within B. longum subspecies.
Pie charts superimposed on the whole genome SNP tree represent predicted GH family gain-loss events within B. longum and B. infantis lineages. Due to the size of the tree, examples of detailed gain loss events have been provided for main lineages, as well as baby BF2 (strains highlighted with light blue) and BF5 (strains highlighted with light purple). Heatmap represents abundance of specific GH families predicted in analysed B. longum strains.

galactotriose from type I arabinogalactans in *B. longum* (*381*). In contrast, strain B_38 isolated from this infant (BF5) post-weaning appears to have lost families GH136 and GH146. Interestingly, enzymes belonging to family GH136 are lacto-*N*-biosidases responsible for liberating lacto-*N*-biose I from LNT, an abundant HMO unique to human milk (*155*), while family GH146 contains β-L-arabinofuranosidases displaying exo-activity on β-linked arabinofuranosyl groups.

These events may possibly be explained by withdrawal of breast milk and/or changes in the composition of the microbiota post-weaning. Only one *B. longum* strain was isolated post-weaning from this baby, however FISH analysis (Figure 3.1 & Table S3.1) revealed an increase in the bacteroides group, which might explain the loss of family GH146 by strain B_38. The founding member of GH146 family, β-L-arabinofuranosidase, was first characterised in *Bacteroides thetaiotaomicron* (*382*). Overall, the presence of intra-individual and strain-specific GH family repertoires in *B. longum* suggests their adaptation to host-specific diet. The presence of strains with different GH content at different dietary stages further indicates potential acquisition of new *Bifidobacterium* strains with nutrient-specific adaptations in response to the changing infant diet.

3.4.6 Prediction of single nucleotide polymorphisms (SNPs) in glycosyl hydrolases

Given the intra-strain diversity in the nine babies and the differences in GH repertoires between *B. longum* and *B. infantis*, I next sought to assess nucleotide-level differences in glycosyl hydrolase genes between strains in individual infants (Table S3.11). Unsurprisingly, I did not identify any significant SNPs that may lead to functional changes in GH genes in infants that had the lowest strain diversity (infants BF1, BF3 and FF2) (Table S3.6). The highest number of GH genes with predicted variants was recorded for *B. infantis* strains from baby BF2. In total, 52 synonymous variants and 29 missense variants were predicted at 81 different positions in 12 GH genes across strains that showed the highest diversity from the first “reference” isolate, namely one strain isolated during weaning and the three strains isolated

post-weaning. A number of missense variants, both complex and single, were recorded at several positions in the predominant enzyme classes, i.e. β-galactosidases (EC 3.2.1.23) and β-N-hexosaminidases (EC 3.2.1.52).

Similarly, both synonymous and missense variants were predicted in *B. longum* strains less closely related to "reference" strains from breast-fed (BF4 and BF5) and formula-fed (FF1, FF3 and FF5) babies. I did not observe any trend in the distribution of SNPs across GH genes in *B. longum* strains. The number of predicted variants, the number of GH genes with identified mutations and their enzyme classification differed between individual infants. For example, in baby BF4 9 out of 10 predicted variants (4 synonymous and 5 missense) were identified in an α-xylosidase in a strain isolated post-weaning, while in baby FF5 14 synonymous and 10 missense variants were predicted at 24 positions in 7 different GH genes across strains isolated during weaning and post-weaning. Some missense changes do not compromise normal protein function, while others can change essential aspects of protein maturation, activity or stability (*383*). The presence of missense variants in GH genes of *B. longum* strains may indicate potential functional differences and provide additional explanation to intra-strain and intra-individual carbohydrate metabolism profiles of these bacteria, however experimental evidence would be essential to confirm the importance of these predictions.

3.4.7 Phenotypic characterisation of carbohydrate utilisation

B. longum strains have previously been shown to be able to utilise a range of carbohydrates, including dietary and host-derived glycans (*195, 365*). Given the predicted differences in carbohydrate metabolism profiles, and to determine strain-specific nutrient preferences, I next sought to assess their glycan utilisation capabilities. I performed growth assays on 49 representative strains from all nine infants, cultured in modified MRS supplemented with selected carbohydrates as the sole carbon source. I chose both plant- and host-derived glycans that I would expect to constitute components of the early life infant diet (*384*). All *B. longum* strains were able to grow on simple carbohydrates (i.e. glucose and lactose). In terms of

growth on more complex carbohydrates, I observed subspecies-specific preferences, consistent with bioinformatic predictions (Figure 3.5).

I selected 2'-fucosyllactose (2'-FL) and lacto-*N*-neotetraose (LNnT) as examples of HMOs found in breast milk, to represent host-derived carbohydrates. Out of the tested isolates, all *B. infantis* strains as well as three *B. longum* strains isolated from a formula-fed baby FF3 during weaning and post-weaning were able to utilise 2'-FL (Figure 3.5). These observations were in line with the results of *in silico* analysis and the prediction of genes potentially involved in degradation of fucosylated carbohydrates in the genomes of these isolates (GH29 and GH95).

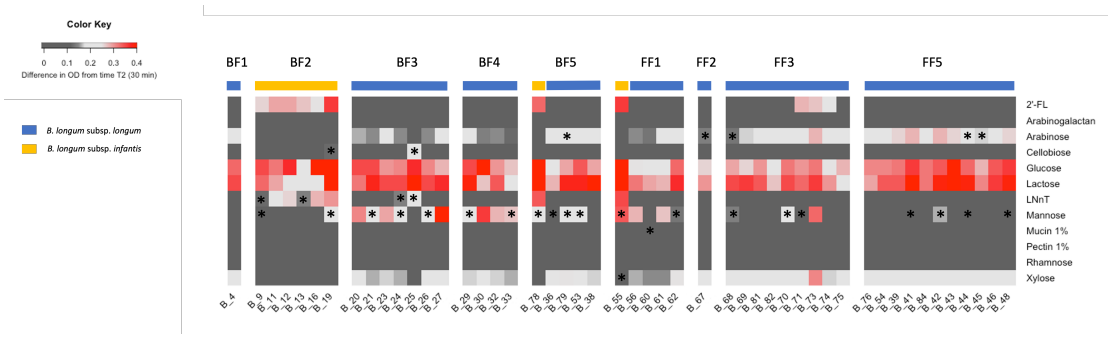


Figure 3.5 Growth performance of B. longum strains isolated from individual infants on different carbon sources.
The coloured bars above the heatmap represent the classification of the isolates into the Bifidobacterium longum subspecies: B. longum – blue, B. infantis – yellow. The heatmap displays the difference in average growth of triplicates between T_2 (30 min) and T_{end} (48 hours). Moderate growth is considered above 0.15 difference in OD from time T_2, high growth above 0.25 difference in OD from time T_2. Asterisks represent strains for which inconsistent growth was recorded (difference in OD of at least 0.15 between any of the duplicates in the triplicate experiment).

Bioinformatic analysis predicted the presence of genes involved in metabolism of isomeric LNT in all *B. longum* strains (GH20 and GH112). However, the ability of *B. infantis* to utilise LNnT was strain-specific, with most strains showing what I considered moderate (above 0.15 difference in OD from time T_2) to high growth rates (above 0.25 difference in OD from time T_2), and two strains displaying inconsistent growth (Table S3.12). Out of *B. longum* strains, B_24 and B_25 (isolated during weaning from breast-fed baby BF3) also grew on LNnT, albeit this was inconsistent. In contrast to all other *B. longum* strains, strain B_25 was not able to

metabolise plant-derived arabinose and xylose despite the predicted presence of genes involved in metabolism of monosaccharides (GH43, GH31, GH2). However, it was one of the two strains (out of 49 tested) that showed growth on cellobiose in two out of the three performed experiments; the other one being the *B. infantis* strain B_19 isolated from baby BF2 post-weaning. Given these interesting results, I performed additional experiments using cellobiose as the sole carbon source over 72h, in which the *B. longum* strain B_25 showed high growth rate (above 0.25 difference in OD from time T_2), while the *B. infantis* B_19 strain did not grow at all (Table S3.12). Furthermore, both *B. longum* and *B. infantis* strains showed varying degrees of growth performance on mannose, even when analysing the same strain, while none of the tested strains were able to grow on arabinogalactan, pectin or rhamnose (Figure 3.5).

To further probe strains identified above for putative carbohydrate degradation genes and gene clusters, I collaborated with Dr Sabina Leanti La Rosa to receive training and perform carbohydrate uptake analysis and proteomics. For these experiments, I chose the breast-fed *B. longum* strain B_25 that showed growth on LNnT and cellobiose, and the formula-fed strain B_71 which was able to grow on 2'-FL. Initially, supernatant from these cultures was analysed using high-performance anion-exchange chromatography (HPAEC) to determine the carbohydrate-depletion profiles (Figure 3.6). In all three cases, the chromatograms showed complete utilisation of the tested carbohydrates and absence of any respective degradation products in the stationary phase culture.

Both cellobiose and 2'-FL were depleted in the early exponential phase by B_25 and by B_71, respectively. LNnT was still detected in the culture supernatant until the late exponential growth phase, which indicated more efficient internalisation of cellobiose and 2'-FL, compared to LNnT. Next, the proteome of B_25 and B_71 when growing on cellobiose, LNnT and 2'-FL compared to glucose was determined (Figure 3.6 a-c & Table S3.13). The top 10 most abundant proteins in the cellobiose proteome of B_25 included three β-glucosidases belonging to GH3 family, as well as a homologue of transport gene cluster previously shown to be upregulated in

B. animalis subsp. *lactis* Bl-04 during growth on this carbohydrate (Figure 3.6 a & Table S3.14) (*385*).

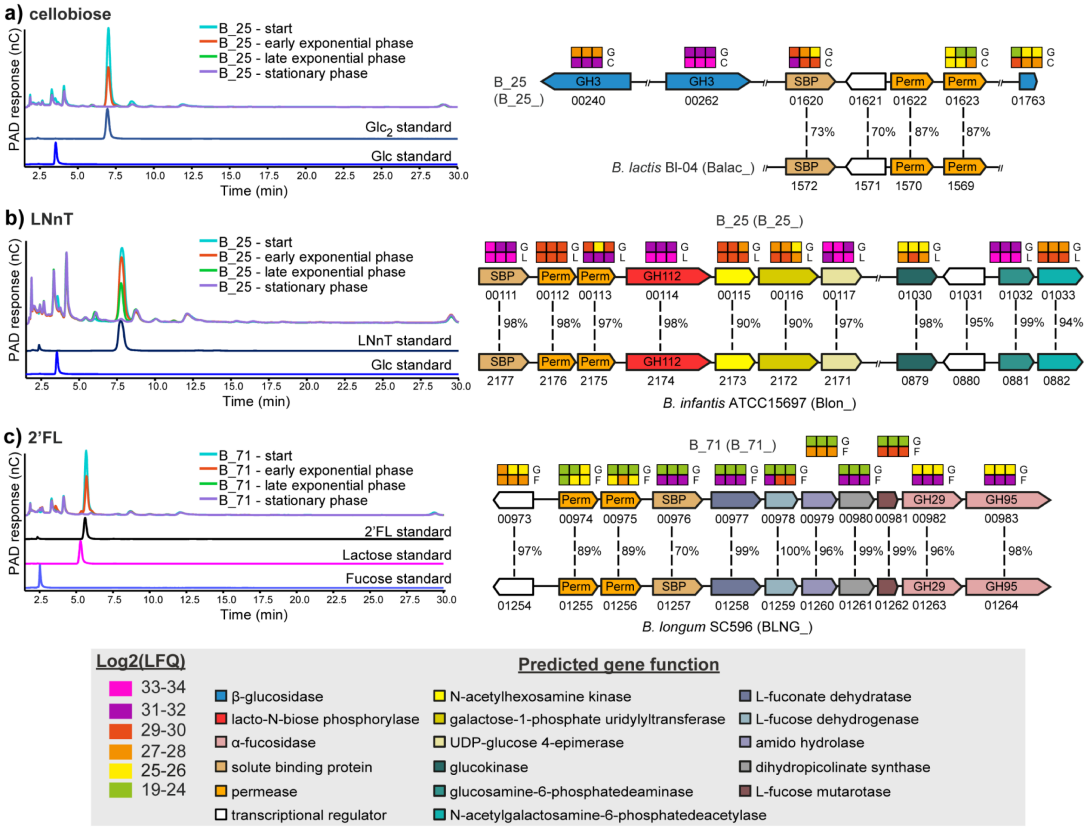


Figure 3.6 HPAEC-PAD traces showing mono-, di- and oligo-saccharides detected in the supernatant of either B_25 or B_71 single cultures during growth in mMRS supplemented with (a) cellobiose; (b) LNnT; (c) 2'-FL.
The data are representative of biological triplicates. Abbreviations: LNnT, Lacto-N-neotetraose; Glc, glucose; Glc2, cellobiose; 2'-FL, 2'-fucosyllactose. Panel on the right shows (a) cellobiose; (b) LNnT; (c) 2'-FL utilization clusters in B_25 and B_71 and proteomic detection of the corresponding proteins during growth on HMOs. Heat maps above genes show the LFQ detection levels for the corresponding proteins in triplicates grown on glucose (G); cellobiose (C); LNnT (L); and 2'-FL (F). Numbers between genes indicate percent identity between corresponding genes in homologous PULs relative to strains B_25 and B_71. Numbers below each gene show the locus tag in the corresponding genome. Locus tag numbers are abbreviated with the last numbers after the second hyphen (for example B_25_XXXXX). The locus tag prefix for each strain is indicated in parenthesis beside the organism name.

Among the three β-glucosidases, B_25_00240 showed 98% sequence identity to the structurally characterized BlBG3 from *B. longum*, which has been shown to be involved in metabolism of the natural glycosides saponins (*386*). B_25_01763 and

B_25_00262 showed 46% identity to the β-glucosidase Bgl3B from *Thermotoga neapolitana* (*387*) and 83% identity to BaBgl3 from *B. adolescentis* ATCC 15703^T (*388*), respectively, two enzymes previously shown to hydrolyse cello-oligosaccharides. With respect to LNnT degradation by the same strain, the most abundant proteins were encoded by genes located in two gene clusters (B_25_00111-00117 and B_25_00130-00133) with functions compatible with LNnT import, degradation to monosaccharides and further metabolism. The identified gene clusters contain the components of an ABC-transporter (B_25_00111-00113), a predicted intracellular GH112 lacto-*N*-biose phosphorylase (B_25_00114), an *N*-acetylhexosamine 1-kinase (B_25_00115) and enzymes involved in the Leloir pathway. All these proteins were close homologues to proteins previously implicated in the degradation of LNT/LNnT by type strain *B. infantis* ATCC 15697^T (*291*) (Figure 3.6 b & Table S3.14). Interestingly, based on the results of bioinformatic predictions, all clonal strains isolated from twin babies BF3 and BF4 contained close homologues of all the above-mentioned genes in their genomes (in some cases identical to those determined in B_25), however, only strain B_25 was able to grow on cellobiose and LNnT. Growth of B_71 on 2'-FL corresponded to increased abundance of proteins encoded by the gene cluster B_71_00973-00983. These proteins showed close homology to proteins described for *B. longum* SC596 and included genes for import of fucosylated oligosaccharides, fucose metabolism and two α-fucosidases belonging to the families GH29 and GH95 (Figure 3.6c & Table S3.14) (*158*).

3.5 Discussion

There is a strong link between high abundance of *Bifidobacterium*, particularly *B. longum*, in early infancy and nutrient availability. The dominance of this species in breast-fed infants has been correlated with enrichment of genes required for the degradation of HMOs present in breast milk, while the transition to solid foods during weaning has been linked to genes involved in degradation of complex plant-derived carbohydrates (*70, 88, 155*). The aim of this study was to investigate the adaptations of *B. longum* to the changing infant diet during the early life

developmental window. I examined wider microbiota composition based on FISH data originally generated by Roger and McCartney (*296*) and assessed the intra-subspecies genomic diversity of 75 *B. longum* strains isolated from nine individual infants at different dietary stages, i.e. pre-weaning, weaning and post-weaning, focussing on their potential carbohydrate utilisation capabilities. In addition, I determined growth performance of representative strains on selected carbohydrates as sole carbon sources. The results suggest intra-individual and diet-related differences in genomic content of analysed strains, which links to their ability to metabolise specific dietary components.

The FISH results are in line with findings from previous studies investigating the infant gut microbiota. I observed inter-individual variability during pre-weaning and weaning, with a shift towards a more adult-like faecal microbiota linked to more complex diet at post-weaning across all samples (*88, 241*). During pre-weaning and weaning, *Bifidobacterium* constituted the predominant group in breast-fed infants, while the composition of microbiota in the formula-fed group during these stages showed higher complexity.

In line with previous reports (*289, 356*), the results of comparative genomic analysis suggest that despite significant changes in diet during weaning, clonal strains of *B. longum* can persist in individuals through infancy, for at least 18 months. At the same time new strains displaying different genomic content and potential carbohydrate metabolism capabilities can be acquired, possibly in response to the changing nutritional environment. Strain shift related to withdrawal of breast milk has previously been suggested for *B. infantis* by Vatanen et al. (*288*) based on a strain-level metagenomic approach. Similarly, work of Asnicar et al. (*389*) indicated that strains of *B. longum* acquired at birth can be replaced at later life stages. Initial vertical acquisition of maternal *Bifidobacterium* strains by newborn babies has been well documented (*146, 368, 389, 390*); however, details of strain acquisition and transmission later in life are currently unclear. Work of Odamaki et al. (*289*) suggested person-to-person horizontal transmission of a particular *B. longum* strain within one family, with direct transfer, common dietary sources or environmental reservoirs, such as family homes (*391*), as potential transmission vehicles and

routes. In this study, I observed the presence of clonal strains in identical twins BF3 and BF4, which may have resulted from a maternal transfer event. However, potential strain transmission between these infants living in the same environment may also occur. Wider studies involving both mothers and twin babies (and other siblings) could provide details on the extent, timing and location of transmission events between members of the same household.

Prediction of genes belonging to GH families was another aspect of performed comparative genomic analysis. The results were in line with previous findings and suggested genome flexibility within *B. longum*, with strains belonging to the same subspecies displaying differences in GH family content. *B. infantis* was found to be predominantly enriched in GH families implicated in the metabolism of host-derived breast milk-associated components like HMOs, while *B. longum* predominantly contained GH families involved in the degradation of plant-derived substrates (*33, 38*). Previously, Vatanen et al. (*288*) suggested that the presence of the HMO gene cluster allowing for intracellular HMO utilisation in *B. infantis* strains confers a particular competitive advantage for this subspecies, leading to its higher relative abundance in the early life microbiota. The results of the analysis of *B. infantis* group indicated the presence of glycosyl hydrolases associated with HMO metabolism in all isolates and revealed subspecies-specific differences in GH content between pre- and post-weaning strains. Furthermore, the data suggested that there are differences in the number of genes belonging to the most abundant GH families (e.g. GH43) between breast-fed and formula-fed strains at different dietary stages, which can be linked to nutrient availability. Surprisingly, both computational and phenotypic approaches identified closely related weaning and post-weaning *B. longum* strains capable of metabolising HMOs (i.e. 2'-FL) in a formula-fed baby that only received standard formula not supplemented with any prebiotics or synthetic HMOs. Moreover, the prediction of SNP variants in genes identified as glycosyl hydrolases indicated the presence of missense mutations in both *B. longum* and *B. infantis* strains. Since some missense variants can compromise protein function (*383*), these results indicate potential functional differences that could further explain intra-strain and intra-individual carbohydrate utilisation properties of

B. longum. However, experimental validation would be essential to confirm the importance of variant predictions.

Recorded phenotypic growth data complemented the results of genomic analyses and further highlighted differences in carbohydrate metabolism profiles between and within *B. longum* and *B. infantis*. As suggested above, the ability of *B. infantis* to utilise different types of HMOs may facilitate their establishment in early life. Similarly, the preference of *B. longum* for plant-based nutrients may be one of the factors influencing their ability to persist within individual hosts through significant dietary changes. The differences in growth of genotypically similar *B. longum* strains on various carbohydrate sources and the ability of formula-fed strains to metabolise selected HMOs suggest that *Bifidobacterium* possess an overall very broad repertoire of carbohydrate utilisation genes that may be differentially switched on and off in response to the presence of specific dietary components (*392, 393*). Potential influence of the intra-individual environment on epigenetic mechanisms in these bacteria may provide another explanation for these results. One potential factor involved in this process may be strain cooperation supported by cross-feeding activities among *Bifidobacterium*, or between *Bifidobacterium* and other members of the microbiota, e.g. *Bacteroides* and *Eubacterium* species (*38, 162, 394, 395*). Indeed, the FISH results revealed the presence of members of these groups, Bac303 (bacteroides) and Erec482 (eubacterium), in faecal samples of both breast- and formula-fed infants, with intra-individual variation at different dietary stages.

Glycan uptake analysis and proteomics resulted in the determination of the mechanisms employed by selected *B. longum* strains to metabolise different carbohydrates. The predicted activity of the most abundant proteins detected during growth on cellobiose, LNnT and 2'-FL, indicated that all these carbohydrates were imported intracellularly and "selfishly" degraded, therefore limiting release of degradation products that could allow cross-feeding by other gut bacteria. These results corroborated the findings from the carbohydrate uptake analysis, where no peak for cellobiose, LNnT and 2'-FL degradation products could be detected. The uptake of cellobiose in B_25 appears to occur via a mechanism similar to that of *B. animalis* subsp. *lactis* Bl-04 (*385*), with cellobiose hydrolysis mediated by the

activity of three intracellular β-glucosidases; however, further confirmatory biochemical characterization of these enzyme is required. B_25 was observed to utilise LNnT using a pathway similar to that previously described in *B. longum* subsp. i*nfantis,* whereby LNnT is internalized via an ABC-transporter (B_25_00111-00113) followed by intracellular degradation into constituent monosaccharides by a GH112 (B_25_00114) and an *N*-acetylhexosamine 1-kinase (B_25_00115). LNnT degradation products are further metabolized to fructose-6-phosphate by activities that include B_25_00116-00117 (galactose-1-phosphate urydyltranferase, UDP-glucose 4-epimerase, involved in the Leloir pathway) and B_25_01030-01033 (for metabolism of *N*-acetylgalactosamine) prior to entering the *Bifidobacterium* genus-specific fructose-6-phosphate phosphoketolase (F6PPK) pathway (*291*). Strain B_71 was predicted to deploy an ABC-transporter (B_71_00974-00976) that allows uptake of intact 2'-FL that is subsequently hydrolysed to L-fucose and lactose by the two predicted intracellular α-fucosidases GH29 (B_71_00982) and GH95 (B_71_00983). L-fucose is further metabolized to L-lactate and pyruvate, via a pathway of non-phosphorylated intermediates that include activities of L-fucose mutarotase (B_71_00981), L-fucose dehydrogenase (B_71_00978), L-fuconate hydrolase (B_71_00977) as previously described for *B. longum* subsp. *longum* SC596 (*158*). Given that the proteins encoded by the aforementioned genes are located in the cellobiose, LNnT and 2'-FL gene clusters that share high similarity and similar organization with those found in equivalent systems in other *B. longum* and *B. animalis*, it is reasonable to suggest that the gene clusters are related and may be the results of horizontal gene transfer events. Collectively, these data reflect inter- and intra-host phenotypic diversity of *B. longum* strains in terms of their carbohydrate degradation capabilities and suggest that intra-individual environment may influence epigenetic mechanisms in *Bifidobacterium*, resulting in differential growth on carbohydrate substrates.

In conclusion, this study provides new insights into distinct genomic and phenotypic abilities of *B. longum* species and strains isolated from the same individuals during the early life developmental window by demonstrating that subspecies- and strain-

specific differences between members of *B. longum* spp. in infant hosts can be correlated to their adaptation at specific age and diet stages.

3.6 Future work

Through the combination of bioinformatic approaches with experimental techniques, this study provided additional insights into genomic and phenotypic features of *B. longum* species and strains isolated from individual infant hosts during the early life developmental window. However, one important limitation was the small number of *B. infantis* strains (n=13) available for analysis, coupled with the fact that the majority of these strains (n=11) were isolated from a single breast-fed baby. Although the investigation of these strains provides important insight into the properties of this subspecies during the transition from breast milk to more diversified diet, it is difficult to assess how representative these results are of the wider population. Furthermore, only one *B. infantis* strain was available from a formula-fed baby, making it impossible to examine properties of members of this subspecies within this dietary group and make comparisons with breast-fed strains. Moreover, only one bacterial strain per time point was available for analysis. Inclusion of additional strains would contribute further observations on inter-individual diversity and functional properties of *Bifidobacterium* in infant hosts.

To assess bacterial communities in faecal samples, I re-analysed the data originally generated using FISH. With detection limit of $\sim 10^6$ bacterial cells (wet weight faeces)$^{-1}$ (*296*), this technique allows investigation of important bacterial groups, but faecal samples may contain a number of organisms at levels below the methodological detection threshold. Furthermore, this technique does not allow for tracking of microbial populations at species level. The use of comprehensive sequencing methods, such as shotgun metagenomics, combined with advanced computational methods to achieve strain-level resolution would allow more detailed examination of bacterial communities in individual hosts.

Phenotypic experiments revealed inconsistencies in growth of individual *B. longum* strains on such carbohydrates as LNnT, cellobiose and mannose. Previously, variability in growth of *B. longum* on mannose, even when analysing the same strain

(*Bifidobacterium longum* NCC2705) has been reported (*396, 397*). Additional carbohydrate metabolism experiments would help determine whether inconsistency in growth on specific carbohydrate sources is inherent to particular *Bifidobacterium* strain. Furthermore, although *B. infantis* is primarily associated with the degradation of HMOs, I recorded growth of one of the *B. infantis* strains from formula-fed baby FF1 on xylose, albeit with inconsistent outcomes between experiments. Further growth assays would help assess the ability of *B. longum* subsp. *infantis* strains to degrade a wider range of non-HMO carbohydrates in early life.

Finally, no metadata on complementary foods during weaning and infant diet post-weaning were available for this study. This information would allow bioinformatic predictions to be related to carbohydrate degradation properties of *B. longum* and the specific dietary components present in weaning infant foods. Future longitudinal studies could be designed to include these data.

Chapter 4

Wild mice are enriched in *Bifidobacterium castoris* strains that circulate within populations and geographical regions and encode specialised genomic signatures related to carbohydrate metabolism and host modulation.

Faecal samples were collected by Dr Sarah Knowles, Miss Aura Raulo and Dr Laima Baltrūnaitė between December 2015 and December 2018.

I carried out all bacterial isolations, DNA extractions and processing for initial identification of the isolates based on the 16S rRNA gene sequences, as well as later WGS sequencing of those isolates initially identified as *Bifidobacterium* spp.

DNA library preparation for WGS was done by sequencing teams at the Wellcome Sanger Institute (Hinxton, UK) and Quadram Institute Bioscience (Norwich, UK).

Biochemical characterisation of strain LH_867 was performed by DSMZ Identification Services (Braunschweig, Germany), except for growth in different pH values, which was carried out by myself.

I performed all the genomic and phylogenomic analyses.

4.1 Introduction

Members of the gut microbiota genus *Bifidobacterium* are widely distributed in the animal kingdom and are believed to exert beneficial effects on their hosts via a variety of mechanisms. In comparison to human-associated strains, in-depth genomic analyses of animal-associated *Bifidobacterium* species and strains are somewhat lacking, particularly in wild animal populations. In this study, I isolated *Bifidobacterium* strains from faecal samples obtained from wild small mammals from two distinct geographical regions in Europe (UK and Lithuania), and performed whole genome sequencing and bioinformatic analysis. The *Bifidobacterium* strains identified (belonging to the species *Bifidobacterium castoris*, *B. animalis* and *B. pseudolongum*) were only found in wild mice (*Apodemus sylvaticus*, *Apodemus agrarius* and *Apodemus flavicollis*) but not voles. Further analysis focusing on the most commonly isolated species (*B. castoris*, 80% isolates) revealed three major mouse-associated clades. These clades were not location- or host species-specific, and their distribution across the host phylogeny was consistent with regular host shifts rather than host-microbe co-speciation. At a finer within-clade level, most *B. castoris* strains were only detected in a single population, but populations frequently harboured multiple co-circulating strains. Functional level analysis indicated that *B. castoris* strains encoded an extensive arsenal of carbohydrate-active enzymes, and a range of glycoside hydrolases including putative novel enzymes such as chitosanases, that may act on chitin-derived substrates such as mushrooms or insects. Gene loss/gain analysis suggested that the *B. castoris* species has acquired the GH49 family, whereas at the strain level, those colonising mice showed mostly glycoside hydrolase gene losses rather than gain events. Other key genomic features uncovered include the presence of putative exopolysaccharides, known to modulate host immune responses, which may have been acquired (alongside carbohydrate metabolism genes) via horizontal gene transfer, as well as the presence of CRISPR-Cas systems, which act as prokaryotic immune defences.

4.2 Background

Species and strains belonging to the bacterial genus *Bifidobacterium* are prominent members of the gut microbiota in many animals, and are universally distributed among animals exhibiting parental care, including humans and non-human mammals, birds and social insects (*6*). *Bifidobacterium* species that colonise the human gut, especially those associated with early life stages, have received much attention in recent years due to their ability to confer positive health benefits on their host, including supporting development of the wider gut microbial ecosystem, colonisation resistance against pathogens and immune modulation. These beneficial properties have been linked to their carbohydrate metabolism and exopolysaccharide (EPS) biosynthesis capabilities (*48, 398*).

Next-generation DNA sequencing, together with comparative and functional genomics, have been shown to be particularly useful in exploring the diversity of *Bifidobacterium* and the genetic basis for their beneficial properties and adaptation to the specific nutritional environment (i.e. host gut). Currently, over 80 *Bifidobacterium* species and subspecies have been identified in multiple animal hosts, with around 2,100 genome assemblies available on the NCBI Genome database (July 2020) (*5*). However, the majority of available genome sequences come from human-associated bifidobacteria (*B. longum*, *B. breve*, *B. bifidum, B. pseudocatenulatum*). Consequently, it is mainly these taxa, as well as type strains of the genus *Bifidobacterium*, that have been subjected to in-depth genomic analysis (*6, 27, 28, 33, 36, 38, 61*).

Recently, work by Lugli et al. (*292, 293*) provided insights into genotypic and phenotypic properties of animal-associated *B. animalis* and *B. pseudolongum* taxa. These species were found in multiple animal hosts, including mammals and birds, and thus appear to be generalists. However, information relating to other animal-associated *Bifidobacterium* species at the genetic variant or strain level is limited, therefore further studies are required to determine if this generalism holds true at higher taxonomic resolution, or whether there is strain-specific host specialisation. This is particularly important when exploring adaption of bacterial symbionts to the

host gut environment in wild populations (rather than model organisms and captive animals).

Recent reports have proposed that the gut microbiota of animals can affect host phenotype in several ways, including through traits related to digestion and energy acquisition and immune development (*399, 400*). These findings suggest that differences in the composition of symbiotic microbial communities may play an important role in host ecology and evolution, and understanding the related distribution and diversification of key bacterial species may provide important insights into microbe-host interactions, particularly in wild populations (*401*). This knowledge may also be important from an animal conservation and management perspective, for example for the maintenance of natural microbiota diversity through conspecific co-habitation or bioaugumentation through microbiota transplantation (*402*), and for determining the potential for animal populations to respond to environmental (e.g. climate) change, especially in non-captive populations in their natural environment (*403*).

Recent reports of congruent phylogenies between mammals and gut commensal taxa have suggested that animal hosts and their symbionts may evolve and speciate together (*404-407*). Amplicon-based approaches identified hundreds of bacterial clades whose approximate divergence times are consistent with those of mammals (*405*). For example, parallel subclade speciation was suggested for members of *Bifidobacteriaceae*, based on their topological and temporal congruency with the phylogenetic tree of four primate species (*Homo sapiens*, *Pan troglodytes*, *Pan paniscus*, and *Gorilla gorilla*) (*404*). These patterns have been linked to convergent acquisition of function by different bacterial phylogenetic clades, with horizontal gene transfer and gene loss proposed as potential mechanisms involved in the process (*408, 409*).

It remains unclear, which processes govern the evolution of mammalian gut microbial symbionts (*410, 411*). Recently, Groussin et al. (*412*) argued that reciprocal and specific functional dependencies between mammalian hosts and single bacterial clades are not strong enough for co-evolution to occur and result in co-speciation events. Instead, the authors proposed that allopatric speciation,

which implies geographic isolation of the host species and subsequent limited symbiont dispersal and diversification, may lead to host-symbiont co-speciation. According to this model, host adaptation to new conditions following an allopatric event results in an altered intestinal environment, to which symbiotic bacteria can quickly adapt (e.g. different glycan composition of plant-derives substrates in herbivore hosts) (*412*). While studies using methods of broad taxonomic resolution (e.g. down to genus or species-level) provide some information on evolutionary relationships between hosts and their gut microbes, higher (strain-level) resolution studies in natural systems remain scarce, yet may reveal cryptic diversity and patterns of host-specificity and host-microbe evolution not previously appreciated.

Wild small mammals provide an ideal study system in which to explore host- gut microbe relationships in more detail, as they are geographically widespread, diverse, easily trapped and possess a rich gut microbiota that, in the case of rodents and especially mice, can be relatively well-understood from studies of laboratory mice. Therefore, wild rodents were chosen as a target host group in which to profile patterns of *Bifidobacterium* diversity, and determine strain-level and functional adaptation to different host species and geographical regions. To this end, I performed *Bifidobacterium* isolation screens in wild mice and voles from multiple populations in two geographically distinct parts of Europe and subsequently investigated a collection of the identified *Bifidobacterium* genomes. Phylogenomic and functional genomic analysis indicated enrichment for *B. castoris* and particular carbohydrate metabolism and host modulatory and defence properties.

4.3 Hypothesis and aims

Whole genome sequences of *Bifidobacterium* isolated from small mammals display particular genomic features related to carbohydrate metabolism, host modulation and defence mechanisms.

Aims:

1) Isolate *Bifidobacterium* from wild small mammal populations, sequence their genomes and select particularly interesting strains for in-depth genomic analysis
2) Determine genomic features of selected strains, with particular focus on traits directly and indirectly linked to carbohydrate metabolism
3) Further explore genomes of selected strains for other interesting properties, including defence systems

4.4 Results

4.4.1 Isolation of *Bifidobacterium* from small mammal faecal samples

Through collaboration with Dr Sarah Knowles and Dr Laima Baltrūnaitė, between April 2017 and December 2018, I obtained and processed a total of 220 faecal samples from 9 species of small mammals (mice, voles and shrews) caught at 14 sites across two European countries – Lithuania and the UK. In the UK, samples were collected from two mouse species (*Apodemus sylvaticus* and *Apodemus flavicollis*) in both Wytham Woods (Oxfordshire, n=78) and Silwood Park (Berkshire, n=14). In Lithuania, 54 samples from mice (*Apodemus agrarius*, *A. flavicollis*), 61 samples from voles (*Microtus agrestis*, *Microtus arvalis, Microtus oeconomus, Myodes glareolus*) and 13 samples from shrews (*Sorex araneus*, *Sorex minutus*, *Neomys fodiens*) were obtained across 12 trapping sites (Table S4.1).

Typically, de Man, Rogosa and Sharpe (MRS) agar supplemented with mupirocin (50mg/l) and L-cysteine (50mg/l) (MRSCM) has been used for *Bifidobacterium* recovery from faeces. However, previous isolation efforts in the lab using this medium produced a number of "false positive" results from animal samples, relating to the fact that different bacterial species displayed colony morphology similar to that of *Bifidobacterium*. Based on these observations and given the fact that only small amounts of faecal material were available from small mammals (~50mg), I decided to carry out isolations using an alternative medium composed of BHI agar supplemented with mupirocin (50mg/l), L-cysteine (50mg/l) and sodium iodoacetate (7.5mg/l) (please also see Chapter 5.4.1 Notes on the isolation of bifidobacterial species from animal gut microbiota).

The isolation experiments resulted in the recovery of 51 *Bifidobacterium* strains from a total of 32 individuals spanning three wild mice species – *A. flavicollis, A. sylvaticus,* and *A. agrarius*, which corresponded to 21.9% of mouse samples screened, including 27.02% *A. flavicollis* samples, 26.31% *A. sylvaticus* samples and 6.9% *A. agrarius* samples, respectively. No *Bifidobacterium* strains were isolated from voles or shrews. The probability of isolating *Bifidobacterium* varied strongly

across host families (Pearson's χ^2(df=2) = 18.98, $P < 0.001$). Whole genome sequencing of recovered strains yielded a mean of 265-fold coverage for strains sequenced on HiSeq (minimum 172-fold, maximum 300-fold) and 225-fold for samples sequenced on MiSeq (minimum 130-fold, maximum 325-fold). One sequence did not assemble correctly and was removed from the dataset. The initial ANI analysis led to the exclusion of further 17 duplicate genomes representing strains from the same individuals sequenced multiple times. In total, I identified 26 strains as a novel *Bifidobacterium* species, 4 strains as *B. animalis* and a further 3 strains as *B. pseudolongum* (Table S4.1).

Given that strains belonging to the novel species constituted the majority of all recovered *Bifidobacterium* strains and were present in individuals from both geographical regions, I selected one strain for characterisation as a type strain and proceeded with the analysis of whole genome sequences (with the plan to publish results in International Journal of Systematic and Evolutionary Microbiology). However, these efforts were stopped following a recent publication that described five novel bifidobacterial species (*413*). My subsequent ANI re-analysis indicated close relatedness (above 95%) between the 26 wild mice strains and the newly characterised type strain *B. castoris* 2020B^T isolated from a captive beaver (*Castor fiber*), ultimately classifying them as *B. castoris*.

4.4.2 Characterisation of strain LH_867 and comparison with type strain *B. castoris* 2020B^T

Initially, prior to the publication of the description of type strain *B. castoris* 2020B^T (*413*), ANI analysis showed that wild mice isolates, representing the putative novel species, displayed sequence identities below 94% to the type strain *B. italicum* Rab10A^T. Based on these results, I decided to proceed with the genome analysis and biochemical characterisation of strain LH_867 isolated from *A. sylvaticus* from Wytham Woods, UK. In order to evaluate the phylogenetic relationship between strain LH_867 and other (at that time, April 2019) recognised *Bifidobacterium*

species and subspecies, I used 16S rRNA, *rpoB, rpoC, groL, dnaJ, clpC, dnaB* and *xpf* genes, as well as whole genome sequences.

Following isolation, all newly recovered isolates were routinely subjected to molecular typing based on the partial 16S rRNA gene sequence and initially identified using the BLASTN algorithm with default settings against the NCBI 16S ribosomal RNA sequence (Bacteria and Archaea) database (*334*). This approach yielded a partial 16S rRNA gene sequence of 987 bp for LH_867, which subsequently showed similarity of 99.80% (over 100% sequence length) with that of *B. choerinum* Su 806 (Figure 4.1).

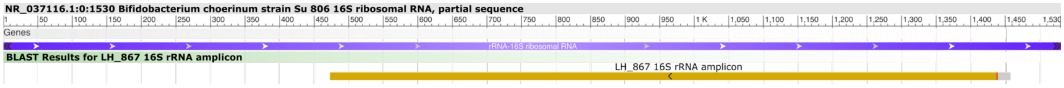


Figure 4.1 Results of the BLASTN similarity search performed for the amplified partial LH_867 16S rRNA gene sequence against the NCBI 16S rRNA gene sequences database, showing similarity over 99% to the 16S rRNA gene of B. choerinum Su 806.

Given that experimental 16S rRNA gene amplification did not yield whole sequence length, the location of 16S rRNA genes was also predicted in the whole genome sequence of strain LH_867 as well as the genomes of type strains and extracted *in silico*. In case of *Bifidobacterium simiarum* TRI 7^T the 16S rRNA gene sequence could not have been predicted, and as it is not available for download from public databases, this strain was excluded from the 16S rRNA gene analysis.

Strain LH_867 showed the highest 16S rRNA gene similarity values with *Bifidobacterium choerinum* LMG 10510^T with 99.41% similarity and *Bifidobacterium italicum* Rab10A^T with 99.12% similarity, respectively (Table 4.1). This result was consistent with the maximum likelihood analysis which indicated phylogenetic relatedness of LH_867 to *B. choerinum* LMG 10510^T (Figure 4.2)

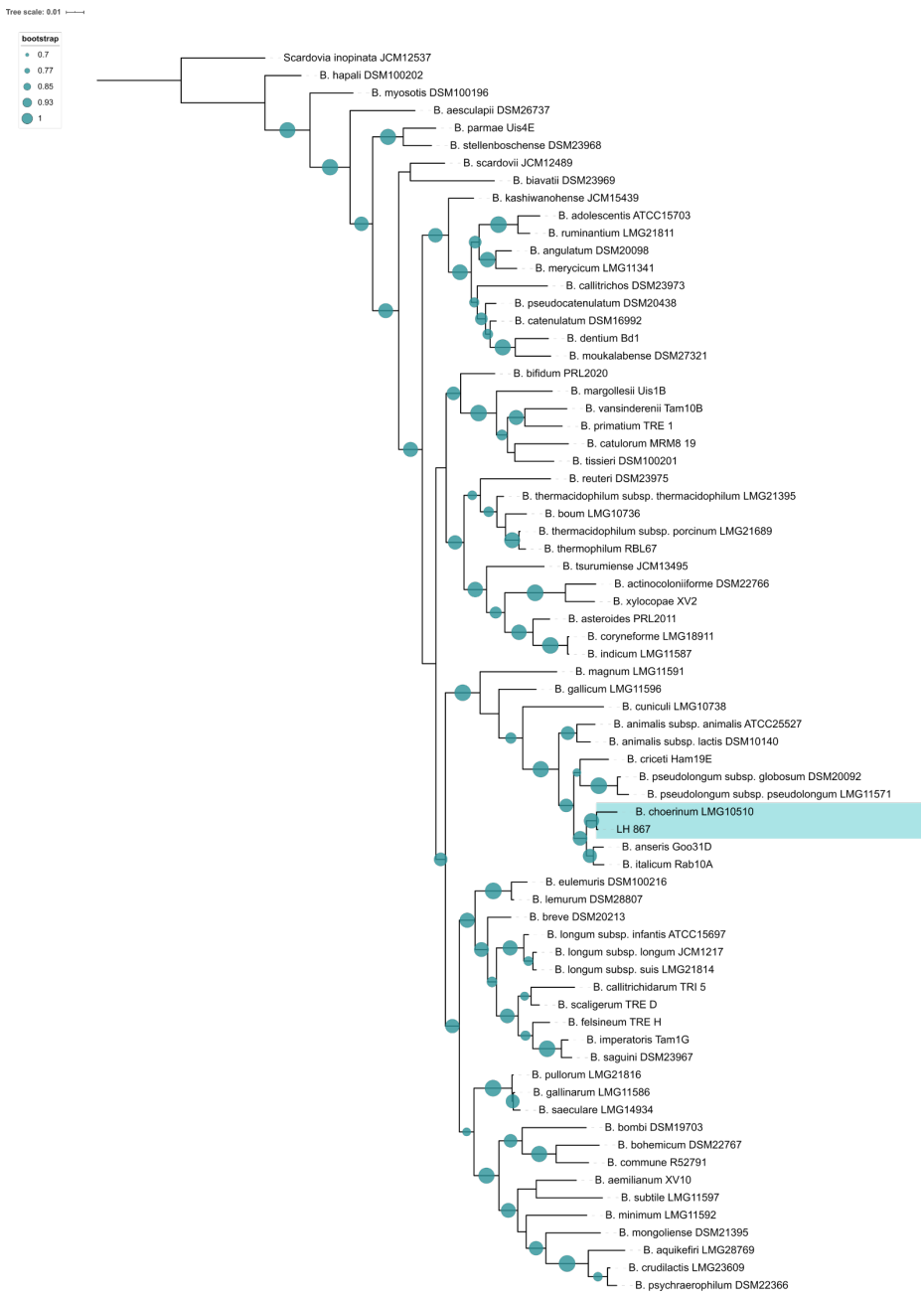


Figure 4.2 Phylogenetic tree based on 16S rRNA gene sequences (1,496 positions) showing relationship of strain LH_867 to type strains of recognised 69 Bifidobacterium species. The tree was reconstructed using the maximum likelihood method with 1000 bootstrap iterations and rooted with Scardovia inopinata JCM 12537^T (=DSM 10107^T). Bootstrap values above 70% are marked on the tree branches.

Previously, multilocus sequence analysis has been proposed as an alternative and robust method for identification and classification of bacterial isolates. Additionally, sequences for housekeeping genes *rpoB, rpoC, groL, dnaJ, clpC, dnaB* and *xpf* genes were *in silico* extracted from all annotated genomes and concatenated for the purpose of multilocus sequence analysis. The resulting multiple sequence

alignment yielded 16,588 positions and was used for the phylogenetic tree reconstruction employing the maximum likelihood method. Based on this analysis, strain LH_867 was found phylogenetically related to *B. italicum* Rab10A^T (Figure 4.3)

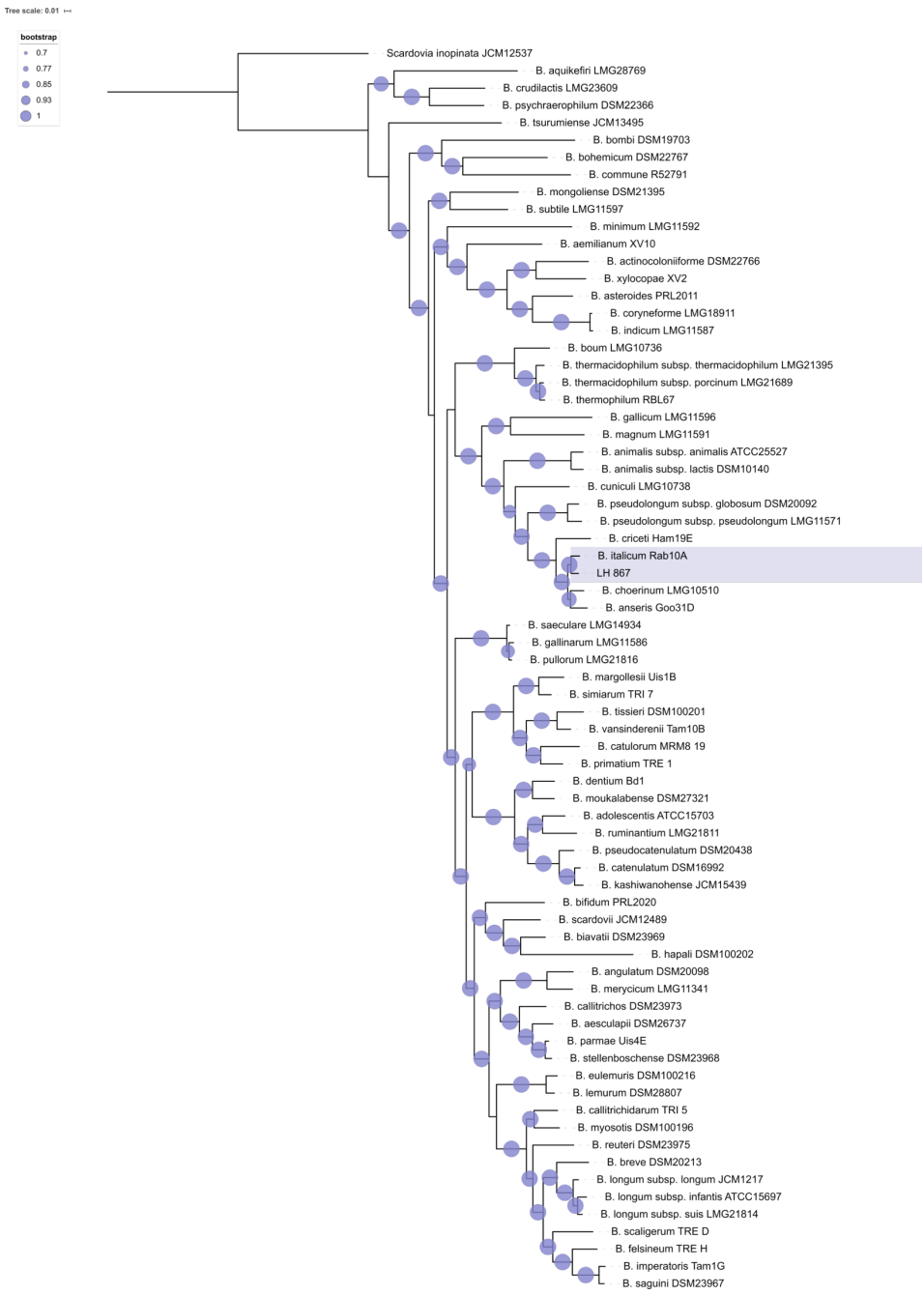


Figure 4.3 Phylogenetic tree based on concatenated housekeeping gene sequences for rpoB, rpoC, groL, dnaJ, clpC, dnaB and xpf genes (16,588 nt) showing relationship of strain LH_867 to type strains of recognised 70 Bifidobacterium species.
The tree was reconstructed using the maximum likelihood method with 1000 bootstrap iterations and the sequence of Scardovia inopinata JCM 12537^T (=DSM 10107^T) was used as an outgroup. Bootstrap values above 70% are marked on the tree branches.

Furthermore, alignment-free composition vector approach was used to reconstruct whole genome-based phylogeny from protein sequences with *k*-mer length of *k*=6. The results of this analysis were consistent with the multilocus sequence analysis and indicated phylogenetic relatedness of LH_867 to *B. italicum* Rab10A^{T} (Figure 4.4 A). Additionally, ANI analysis showed that LH_867 displayed the highest sequence identity to *B. italicum* Rab10A^{T}, with the value of 92.98% (Figure 4.4 B & Table S4.2).

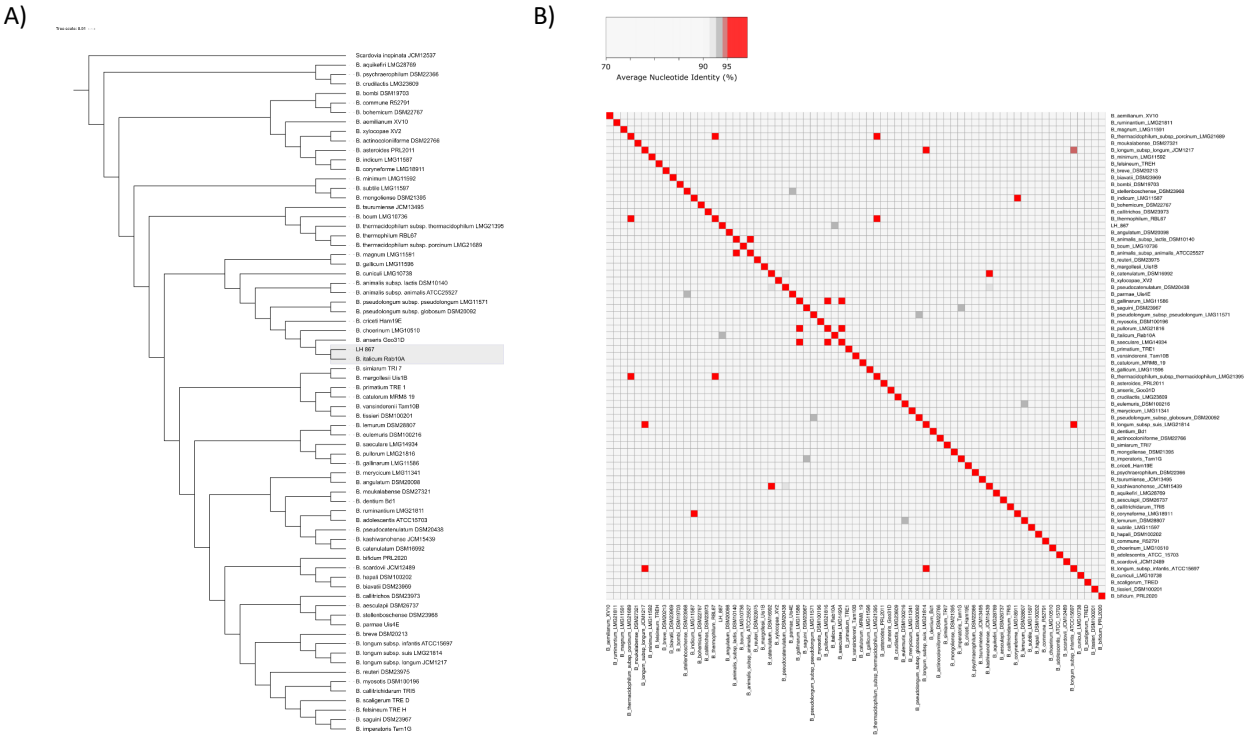


Figure 4.4 Relatedness of LH_867 to type strains of recognised 70 Bifidobacterium species based on whole genome sequences.
(A) Alignment-free whole genome sequence phylogeny constructed based on amino acid sequences with k-mer length of k=6. The sequence of Scardovia inopinata JCM 12537^{T} (=DSM 10107^{T}) was used as an outgroup; (B) Average Nucleotide Identity values between LH_867 and genomes of type strain Bifidobacterium. The diagram shows values between 70-100%, values above 95% are marked in red.

Upon publication of the description of several novel *Bifidobacterium* species (*413*), ANI re-analysis revealed the identity value of 98.69% between the type strain *B. castoris* 2020B^{T} and strain LH_867. Similar results were obtained for the

remaining 25 wild mice isolates, subsequently resulting in their classification as members of this species (Table S4.1). The comparison between *B. castoris* 2020B^T and LH_867 indicated strain specific differences in genomic and phenotypic features, including genome size, number of ORFs and tRNAs, growth conditions and fermentation profiles (Table 4.1 & Table 4.2).

	***B. castoris* 2020B^T**	**LH_867**
Biological origin	*Castor fiber*	*Apodemus sylvaticus*
Geographical location	Italy (captive)	United Kingdom (wild)
Average coverage	195	286
No of contigs	22	15
Genome length	2496067	2280630
Average GC %	65.41	65.48
No of ORFs	2053	1840
tRNA	55	53
16S rRNA gene similarity	98.76% *B. choerinum*	99.41% *B. choerinum*
ANI value	93.8% *B. italicum*	92.98% *B. italicum*

Table 4.1 General genomic features of B. castoris 2020B^T and LH_867.

Optimal growth conditions of both *B. castoris* 2020B^T and LH_867 were determined in MRS after incubation at different temperatures and at different pH in anaerobic conditions for 48 hours. *B. castoris* 2020B^T was reported to grow at the temperature range between 25 to 40 ^oC, but not below or above these values (*413*). In contrast, data for LH_867 showed a wider growth temperature range, from 20 to 45 ^oC. In terms of growth at different pH, *B. castoris* 2020B^T was shown to grow at pH 5-6, however growth at higher pH values was not tested. I observed growth of strain LH_867 at pH ranging from 5 to 8. With regard to enzymatic activity, both strains are catalase and gelatinase negative. Fermentation profiles of *B. castoris* 2020B^T and LH_867 revealed that the strains are able to digest a wide range of simple and complex carbohydrates, and that their carbohydrate metabolism profiles are strain-specific, with recorded differences in growth on substrates including D-ribose, D-galactose, D-mannose, cellobiose, sucrose, raffinose and L-fucose (Table 4.2).

Members of *Bifidobacterium* and related bacterial genera can convert monosaccharides resulting from the degradation of complex carbohydrates to intermediates of a particular hexose fermentation pathway, the ‘‘bifid shunt’’, and ultimately produce SCFAs, e.g. acetate, formate, and other organic compounds (*4, 42, 48*). The key enzyme in the “bifid shunt” is fructose-6-phosphate phosphoketolase (EC 4.1.2.2), considered a taxonomic marker for identification of *Bifidobacteriaceae* (*49*). The results of the prediction of the “bifid shunt” in the genome of strain LH_867 were in line with those reported for *B. castoris* 2020B^T and indicated the presence of the complete hexose fermentation pathway (Table S4.3).

Characteristic	***B. castoris* 2020B^T**	**LH_867**
Carbohydrate metabolism:	1% of substrate in mMRS	API 50 CHL
Glycerol	n/a	-
Erithritol	n/a	-
D-Arabinose	-	-
L-Arabinose	n/a	W
D-Ribose	+	W
D-Xylose	W	W
L-Xylose	n/a	-
D-Adonitol	n/a	-
Methyl α-D-xylopyranoside	n/a	-
D-Galactose	+	W
D-Glucose	+	?
F-Fructose	-	-
D-Mannose	+	-
L-Sorbose	n/a	-
L-Rhamnose	-	-
Dulcitol	n/a	-
Inositol	n/a	-
D-Mannitol	-	-
D-Sorbitol	W	-
Methyl α-D-mannopyranoside	n/a	-
Methyl α-D-glucopyranoside	n/a	-
N-acetylglucosamine	W	-
Amygdalin	n/a	-
Arbutin	n/a	-
Aesculin ferric citrate	n/a	-
Salicin	n/a	-
Cellobiose	+	-
Maltose	n/a	?

Lactose (bovine origin)	n/a	?
Melibiose	W	W
Sucrose	+	W
Trehalose	W	-
Inulin	W	-
Melezitose	n/a	-
Raffinose	+	W
Starch	W	W
Glycogen	n/a	W
Xylitol	n/a	-
Gentiobiose	n/a	W
Turanose	n/a	-
D-Lyxose	n/a	-
D-Tagatose	n/a	-
D-Fucose	n/a	-
L-Fucose	W	-
D-Arabitol	n/a	-
L-Arabitol	n/a	-
Potassium gluconate	n/a	-
Potassium 2-ketogluconate	n/a	-
Potassium 5-ketogluconate	n/a	W
Enzymatic activities:		
Catalase	-	-
Oxidase	n/a	-
Haemolysis	n/a	-
Gelatinase	-	-
Temperature for growth (°C) range	25-40 °C	20-45 °C
pH for growth range	5-6 (tested up to pH 6)	5-8 (tested up to pH 8)
DNA G+C% content	65.41 mol% *(in silico* prediction)	65.0 mol%
Peptidoglycan type	n/a	L-Lys(L-Orn)-L-Ala(L-Ser)-L-Ala_2

Table 4.2 Differential phenotypic characteristics of type strain B. castoris 2020B^T and LH_867. Values for B. castoris 2020B^T adapted from Duranti et al. (2019) (413).

The estimation of G+C% content of LH_867 DNA was performed by the DSMZ Identification Service, Braunschweig, Germany using the Agilent 1260 Infinity II HPLC system. Strain LH_867 had a DNA G+C% content of 65.0 mol%. This value is similar to that reported for strain *B. castoris* 2020B^T (*in silico* prediction) and within the range reported for the genus *Bifidobacterium*: 52–67 mol% (*1*).

The peptidoglycan composition of strain LH_867 was examined by the DSMZ according to protocol used in Schumann, 2011 (*414*). The total hydrolysate (4N HCl,

16h, 100 °C) of the peptidoglycan contained muramic acid (Mur) and the amino acids alanine (Ala), serine (Ser), ornithine (Orn), lysine (Lys) and glutamic acid (Glu) in the following molar ratio: 3.4 Ala, 0.3 Ser, 0.2 Orn, 0.7 Lys, 1.0 Glu and 1.7 Mur. Enantiomeric analysis of the peptidoglycan amino acids revealed the presence of D-Ala and L-Ala in a ratio of 1:3.9. The partial hydrolysate (4N HCl, 0.75h, 100°C) of the peptidoglycan contained (in addition to the amino acids) the peptides Lys-Ala, Lys-$(Ala)_2$ (3 isomers), Lys-$(Ala)_3$ (4 isomers), Orn-Ser, Lys-$(Ala)_4$, Lys-$(Ala)_4$-Lys, M-Ala, Ala-Glu. Based on these results, it was concluded that strain LH_867 represents the peptidoglycan type A3α L-Lys(L-Orn)-L-Ala(L-Ser)-L-Ala_2 (type A11.21 according to www.peptidoglycan-types.info). The peptidoglycan composition of strain *B. castoris* 2020B^T was not reported.

4.4.3 Genomic characterisation of *B. castoris* taxon

Following the classification of the mice strains as *B. castoris*, I performed genomic re-analysis of the entire dataset with strain *B. castoris* 2020B^T included as reference (Table S4.1). The assembled draft genome sizes for the mouse-associated *B. castoris* isolates ranged from 2.27 Mb to 2.39 MB, possessing an average G+C% content of 65.53% and a number of contigs ranging from 9 to 56. The number of predicted ORFs in each genome ranged from 1,832 to 1,980. Genome sizes and gene numbers were therefore lower in comparison to the type strain *B. castoris* 2020B^T, whose genome size is 2.50 Mb, with 2,053 ORFs and an average G+C% content of 65.41% (Table 4.1) (*413*). The sizes of draft genomes for *B. animalis* and *B. pseudolongum* ranged from 2.15 Mb to 2.19 Mb (1,808 to 1,849 ORFs) and 2.03 Mb to 2.06 Mb (1,705 to 1,725 ORFs), respectively. *B. animalis* strains had an average G+C% content of 60.00%, while this value was at 63.34% for *B. pseudolongum*. These findings are in line with previous reports for members of these species isolated from rodents (*292, 293*).

In terms of *Bifidobacterium* distribution across the host species, *A. sylvaticus* (n= 19, UK) was found to harbour *B. castoris* and *B. animalis* strains. *A. flavicollis* in the UK (n=2) only harboured *B. animalis* strains, whereas the same host species in

Lithuania (n=8) harboured both *B. castoris* and *B. pseudolongum* strains. *A. agrarius* (n=2, Lithuania) was found to only harbour *B. castoris* strains. With strains displaying ANI value above 99.9% considered identical (*309*), 5 *B. castoris* and 2 *B. animalis* strains were identified in *A. sylvaticus* population, 5 *B. castoris*, 2 *B. animalis* and 3 *B. pseudolongum* strains were found in *A. flavicollis*, while 2 *B. castoris* strains were present in *A. agrarius*. On average, I recovered one unduplicated *Bifidobacterium* strain per individual, except for one *A. sylvaticus* individual from Wytham (X0418EBC072), who was found to have harboured both *B. castoris* and *B. animalis*.

Interestingly, all newly isolated strains belonged to the previously established *B. pseudolongum* phylogenetic group (*286*). Recent taxonomic analyses of genus *Bifidobacterium* indicated that this phylogenetic group was the most diverse in terms of ecological niches represented by host species, and encompassed strains isolated from animals such as chickens, geese, dogs, oxen, pigs, rabbits, hamsters and rats (*286*). Since the relatedness of organisms can effectively be predicted based on their shared gene content (*415, 416*), I constructed a maximum-likelihood phylogenetic tree using single copy core genes to assess relationships between the wild mice strains and the representative members of the *B. pseudolongum* group (n=112), with particular interest in strains isolated from rodents (Figure 4.5). Despite a limited number of rodent-associated *Bifidobacterium* genomes available for this analysis (n=18), the results show that while there is some clustering of strains according to the host phylogeny, this is far from perfect. For example, while *B. animalis* strains isolated from mice tend to cluster separately from those from rats, the type strain of *B. castoris* from a beaver falls in the middle of a clade otherwise comprised entirely of mouse-isolated strains. This observation is in line with previous analysis of animal-associated *B. pseudolongum* strains, which indicated that different animal hosts may harbour specific clusters of members of this taxon (*293*).

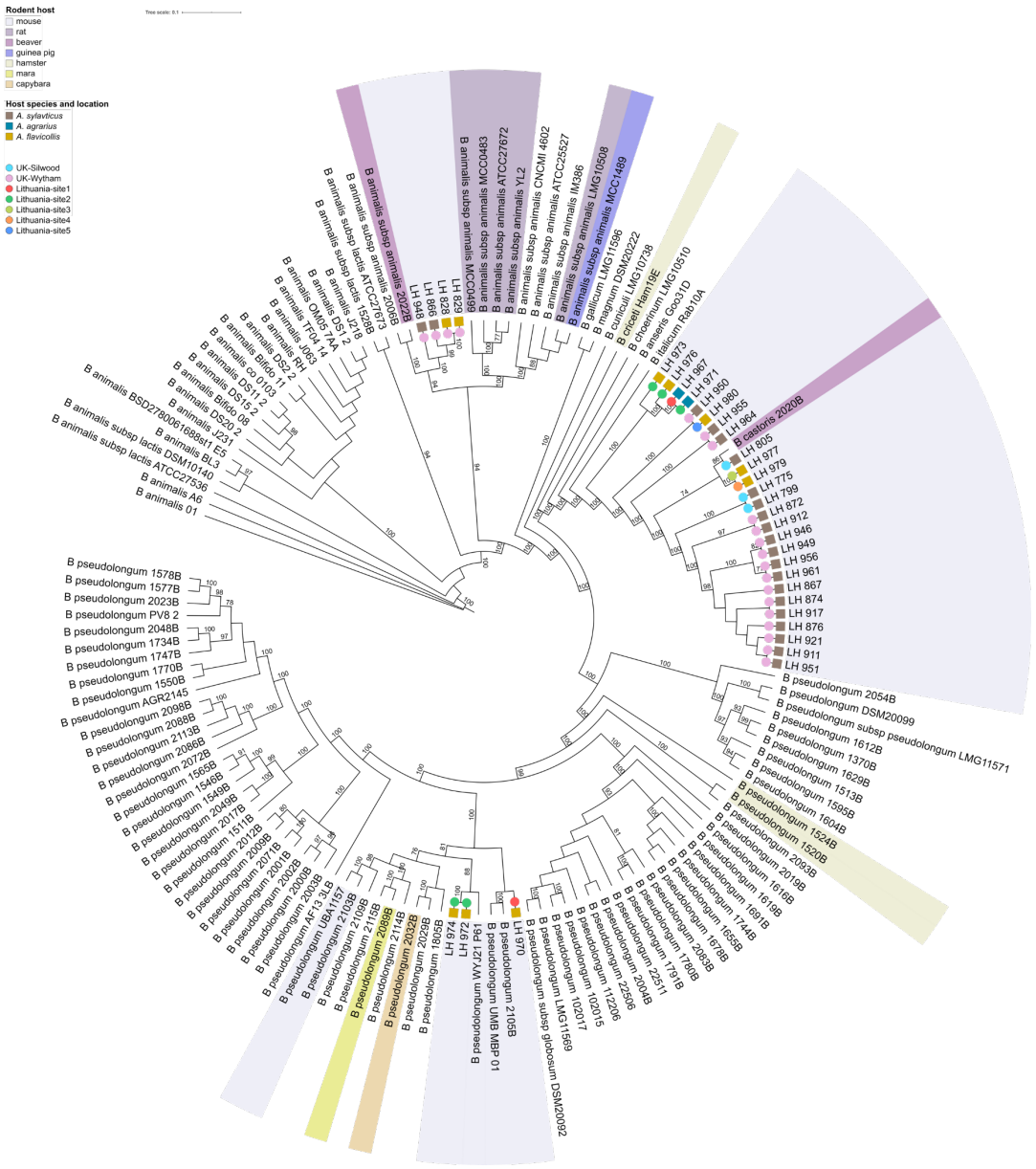


Figure 4.5 Cladogram of Bifidobacterium pseudolongum phylogenetic group, including 112 publicly available representative strains and the 33 strains recovered in this study.
Maximum likelihood phylogeny was based on single copy core genes, employing the 'WAG' general matrix model with 1000 bootstrap iterations. Bootstrap values above 70% are displayed on tree branches. Strains isolated from rodent hosts are marked with coloured background. Coloured symbols on the branches depict respective host species (square) and trapping sites (circle) for strains recovered in this study.

Since *B. castoris* constituted 78% of all recovered *Bifidobacterium* strains recovered in this study and this species is the least well-characterised, I focused further analyses on this species. The pangenomic analysis of 27 strains of *B. castoris* (Figure 4.6) revealed a total of 2,897 gene clusters. Based on the distribution of

gene clusters in the pan-genome, I identified 1,412 gene clusters that constituted the core genome shared by all strains (48.7% of all clusters), while 438 clusters (15.1% of all clusters) were unique genes (Table S4.4). Using protein sequences for single copy genes of the pan-genome, I constructed a *B. castoris* phylogeny based on maximum likelihood estimation (Figure 4.6 and Figure 4.7 A).

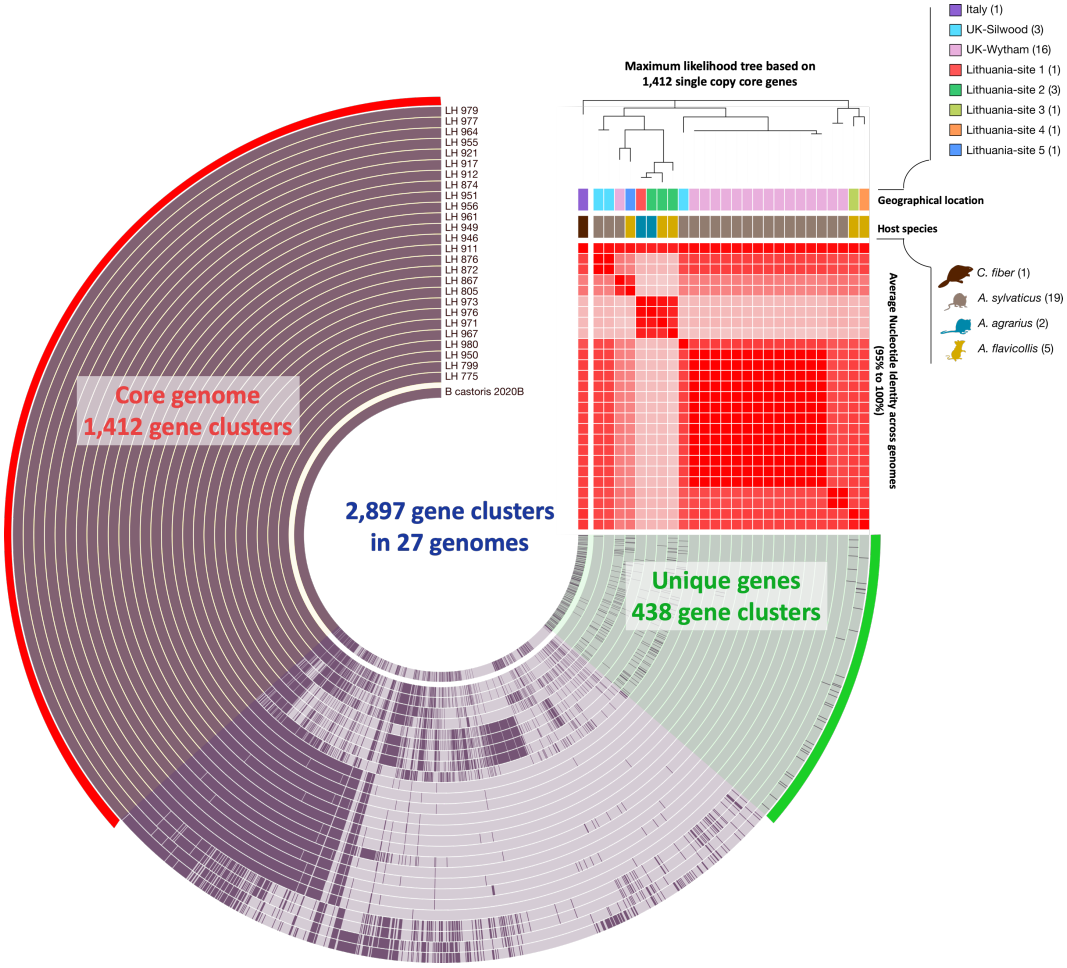


Figure 4.6 Pan-genomic analysis of 27 genomes of B. castoris.
The results revealed 1,412 (48.7% of all clusters) core gene clusters, and 438 (15.1%) strain-specific unique gene clusters among 2,897 total gene clusters. The heatmap represents average nucleotide identities between genomes (ANI > 95%).

Examination of the *B. castoris* phylogenetic tree (Figure 4.7 A), indicated the presence of three major mouse-associated *B. castoris* clusters. One cluster is more distant from the other two and contains strains from both *A. sylvaticus* and *A. flavicollis*. The second cluster seems to be *A. sylvaticus*-specific and appear to contain 2 main strains, while the third main cluster contains strains isolated from all

three *Apodemus* species, including one *A. sylvaticus*-specific subclade, and another subclade that contains two clusters of strains that each colonise two different host species.

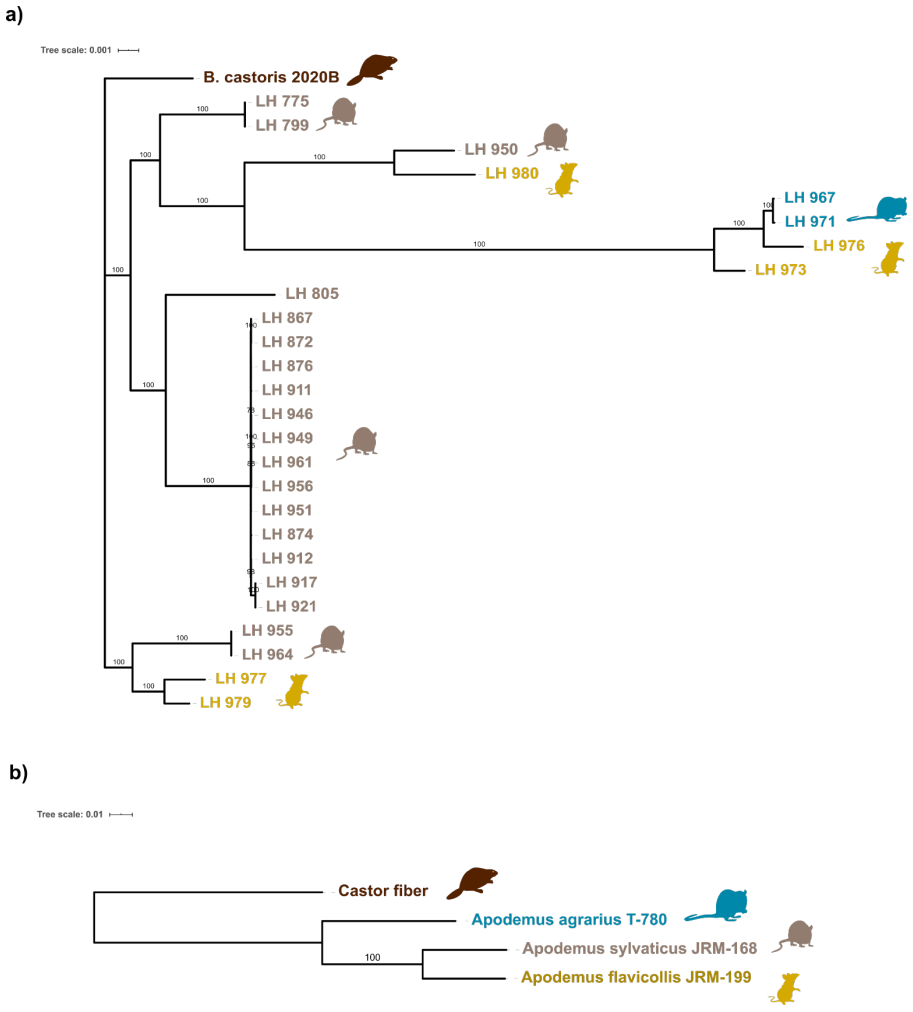


Figure 4.7 Phylogeny of 27 Bifidobacterium castoris strains (A) and their rodent hosts (B). Maximum likelihood trees were constructed using single copy core genes employing WAG model and 1000 bootstrap iterations for B. castoris and concatenated 12S rRNA and partial cytochrome b genes employing 'GTR' model with 1000 bootstrap iterations for host species. Bootstrap values above 70% are displayed on tree branches.

Overall, these observations are not consistent with a strong pattern of co-speciation. Therefore, I sought to test for co-phylogenetic signal between the host and *Bifidobacterium* by performing a topology-based comparison (*329*). The host tree was constructed using concatenated sequences for part of the cytochrome b

(cytb) gene and the mitochondrial 12S rRNA gene (Figure 4.7 B) (*417*). Co-phylogeny was first tested using the global ParaFit statistic (H0: the hosts and *B. castoris* have independent phylogenetic structure). The result (permutational *P* = 0.3588 after 9999 permutations) was not significant, falling to reject the null hypothesis, providing no evidence that the phylogeny of *B. castoris* and its hosts are correlated. Further test of individual host-*Bifidobacterium* associations did not reveal any significant links (*P* > 0.05), suggesting the absence of co-phylogenetic patterns (Table 4.3).

Global test: ParaFitGlobal = 8.835461e-05 , p-value = 0.3588 (9999 permutations)

Test of individual links (9999 permutations)

Link	Host	*B. castoris* strain	F1 stat	F1 p-value	F2 stat	F2 p-value
[26,]	A. agrarius	LH_967	2.66E-05	0.3085	9.07E-03	0.3081
[27,]	A. agrarius	LH_971	2.66E-05	0.3054	9.08E-03	0.3051
[21,]	A. flavicollis	LH_980	-6.22E-07	0.7777	-2.12E-04	0.7777
[22,]	A. flavicollis	LH_976	-1.76E-05	0.8014	-5.99E-03	0.8032
[23,]	A. flavicollis	LH_973	-1.52E-05	0.7789	-5.17E-03	0.7811
[24,]	A. flavicollis	LH_977	-2.70E-08	0.5797	-9.20E-06	0.5797
[25,]	A. flavicollis	LH_979	7.89E-09	0.5611	2.69E-06	0.5611
[10,]	A. sylvaticus	LH_946	6.99E-06	0.2167	2.38E-03	0.2168
[11,]	A. sylvaticus	LH_949	6.99E-06	0.2119	2.38E-03	0.2127
[12,]	A. sylvaticus	LH_961	6.99E-06	0.2195	2.38E-03	0.2193
[13,]	A. sylvaticus	LH_956	6.99E-06	0.2128	2.38E-03	0.2128
[14,]	A. sylvaticus	LH_951	6.99E-06	0.2175	2.38E-03	0.2168
[15,]	A. sylvaticus	LH_874	7.00E-06	0.2117	2.39E-03	0.2115
[16,]	A. sylvaticus	LH_912	6.99E-06	0.2173	2.38E-03	0.2177
[17,]	A. sylvaticus	LH_917	7.21E-06	0.2087	2.46E-03	0.2092
[18,]	A. sylvaticus	LH_921	7.21E-06	0.2158	2.46E-03	0.2157
[19,]	A. sylvaticus	LH_955	3.41E-06	0.4103	1.16E-03	0.4083
[2,]	A. sylvaticus	LH_775	1.14E-06	0.4466	3.90E-04	0.4461
[20,]	A. sylvaticus	LH_964	3.41E-06	0.4081	1.16E-03	0.4067
[3,]	A. sylvaticus	LH_799	1.14E-06	0.4403	3.90E-04	0.4386
[4,]	A. sylvaticus	LH_950	-3.04E-06	0.839	-1.04E-03	0.8392
[5,]	A. sylvaticus	LH_805	6.00E-06	0.279	2.04E-03	0.2777
[6,]	A. sylvaticus	LH_867	6.98E-06	0.2161	2.38E-03	0.2169
[7,]	A. sylvaticus	LH_872	6.99E-06	0.2179	2.38E-03	0.2188
[8,]	A. sylvaticus	LH_876	6.98E-06	0.218	2.38E-03	0.2175
[9,]	A. sylvaticus	LH_911	6.99E-06	0.2135	2.38E-03	0.2135
[1,]	B. castoris	B_castoris_2020B	2.75E-06	0.3883	9.36E-04	0.3854

Table 4.3 Results of co-phylogenetic analysis using the ParaFit statistic with 9999 permutations.

Next, I sought to determine how many different strains circulate in each host population and country. Average identity value for the two strains isolated from

A. agrarius trapped at two different sites in Lithuania (LH_967 and LH_971) was 99.88% (Table S4.5). Overall, strains isolated from *A. flavicollis* had the broadest identity value range (97.07±0.01% mean±sd), suggesting that they are less closely related to each other than strains isolated from *A. agrarius*, as well as *A. sylvaticus* in the UK (99.11±0.01%, mean±sd). Based on identity values between individual isolates in the latter population, and with the majority of strains at Wytham being identical (ANI values > 99.99%, it appears that a total of 5 *B. castoris* strains circulate across the two UK sites: two strains at Silwood, and three at Wytham (with one strain being more common than the others).

4.4.4 Glycobiome of *B. castoris*

To determine functional differences between *B. castoris* strains, I next assigned functional categories to ORFs of each genome. This analysis reflected the saccharolytic lifestyle of this species, with carbohydrate transport and metabolism identified as second most abundant COG category (after unknown function) constituting 9.98% of its pan-genome (Table S4.6). This value is slightly higher compared to previous findings for the pan-genome of the animal-associated *B. pseudolongum* taxon (9%) and within the range reported for other bifidobacteria (*162, 293*). Members of *Bifidobacterium* have been shown to synthesize and digest a wide range of carbohydrates through an extensive arsenal of carbohydrate-active enzymes (CAZymes)(*42, 258*). I thus sought to investigate the genetic repertoire predicted to be involved in carbohydrate metabolism and biosynthesis in *B. castoris*. *In silico* analyses performed using dbCAN2 identified three classes of enzymes, namely glycoside hydrolases (GHs), glycoside transferases (GTs) and carbohydrate esterases (CEs), as well as enzyme-associated carbohydrate-binding modules (CBMs) (Figure 4.8 & Table S4.7). On average, *B. castoris* genomes harboured 87.52±3.15 CAZymes. Previous reports on CAZyme abundances in strains isolated from different hosts and environments showed that, on average, *Bifidobacterium* isolated from rodents had less than 50 CAZyme genes in their genomes, a number comparable with strains isolated from dairy and wastewater

(*31*). Interestingly, the abundance of CAZymes similar to that of *B. castoris* was reported for strains isolated from non-human primates (84±10 CAZymes) (*31*).

Glycosyl hydrolases are key enzymes in carbohydrate metabolism that catalyse the hydrolysis of glycosidic bonds between two or more carbohydrates or between a carbohydrate and non-carbohydrate moiety (*418*). I identified a total of 25 different GH families in *B. castoris* strains containing an average of 50.48±1.99 GH genes per genome (2.67%±0.16 of ORFs and 57.68±0.01% of predicted glycobiome). The predominant GH family, with 15.15±1.29 GH genes per genome (mean±sd), was GH13, whose members include enzymes acting on a very wide range of carbohydrates containing α-glucoside linkages, e.g. starches and related substrates, trehalose, raffinose, stachyose and melibiose (*38, 42, 419*). This was followed by families GH31 (a diverse group of enzymes with α-glucosidase and α-xylosidase activities) and GH36 (enzymes metabolising α-galacto-oligosaccharides present in various plants, i.e. melibiose, raffinose, stachyose) (*420, 421*), with 3.63±0.68 and 3.18±0.39 GH genes per genome, respectively.

Strain-specific predictions of GH repertoires in *B. castoris* seem to mostly support previous observations on strains circulating in *Apodemus* populations. In contrast with the majority of strains isolated from *A. sylvaticus* in the UK, strains from *A. flavicollis* in Lithuania displayed individual differences in GH profiles, with a subset harbouring genes predicted to encode members of GH46 containing chitosanases acting on chitin-derived substrates (i.e. mushrooms, fungi and insects), and GH127 encompassing enzymes with β-L-arabinofuranosidase activity (Figure 4.8 & Table S4.7) (*422, 423*). In the UK, 13 out of 16 isolates from *A. sylvaticus* from Wytham appeared to have identical numbers of genes belonging to specific GH families, while the remaining 3 isolates displayed different GH profiles. These results support previous findings that there appear to be 4 strains circulating in the *A. sylvaticus* population at this location. However, I did not observe this consistency with isolates from *A. sylvaticus* from Silwood. The two isolates which showed ANI value of 99.99% (LH_775 and LH_799) displayed differences in the number of predicted GH13 genes (14 vs 16), with identical values for the other identified GH families. This result may possibly be explained by the gene calling process during the

annotation of draft assemblies and the resulting differences in the number of predicted ORFs (1,942 for LH_775 vs 1,933 for LH_799, Table_S4.1).

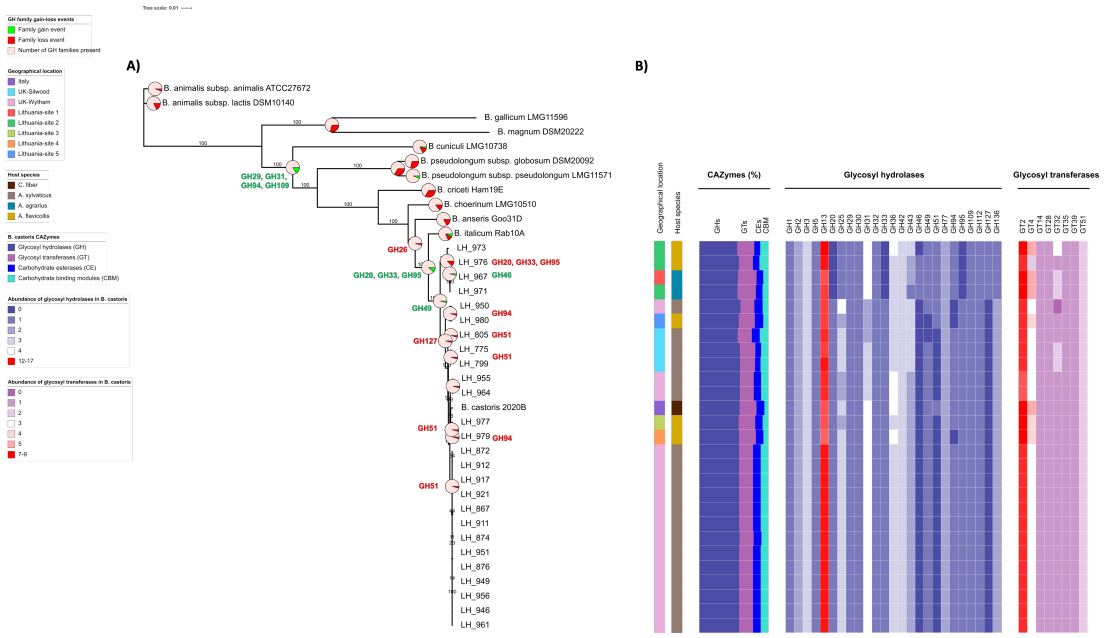


Figure 4.8 Glycosyl hydrolase (GH) family gain-loss events in B. castoris and the type strains representative of the B. pseudolongum phylogenetic group (A), and the abundance of carbohydrate-active enzymes (CAZymes) in B. castoris (B) .

(A) Maximum likelihood phylogeny reconstruction was based on single copy core genes of the B. castoris pan-genome, employing the 'WAG' general matrix model with 1000 bootstrap iterations. Dollo parsimony was used to predict GH family gain-loss events.

(B) The coloured strips represent isolation period geographical locations and host species, respectively. The bar plot shows proportional representation of four CAZyme classes constituting the predicted B. castoris glycobiome. The two heatmaps present abundances of genes predicted to belong to specific glycosyl hydrolase (GH) and glycosyl transferase (GT) families.

Glycosyl transferase class of enzymes catalyse the formation of glycosides involved in the biosynthesis of oligosaccharides, polysaccharides and glycoconjugates (*424*) and have previously been associated with production of exopolysaccharide (EPS) in different bacterial species (*425*). A total of 8 GT families were predicted in the *B. castoris* genomes, with 18.70±1.61 GT genes on average (21.41±0.02% of predicted glycobiome). GT2 family was predominant in all analysed strains, with an average of 8.15±0.53 GT genes per genome. Carbohydrate esterases, whose function is to release acyl or alkyl groups attached by ester linkage to carbohydrates

(*426*), and carbohydrate-binding modules, which have no hydrolytic activity, but bind to carbohydrate ligands and enhance the catalytic efficiency of carbohydrate active enzymes (*426*), constituted 10.66±0.01% and 10.24±0.02% of predicted glycobiome, with 9.33±1.00 and 9.00±1.75 genes per genome, respectively (Figure 4.8 & Table S4.7). Overall, these findings highlight a predominance of genes encoding GH families predicted to be responsible for the breakdown of plant-derived polysaccharides in the genomes of *B. castoris* species.

4.4.5 Glycosyl hydrolase gene gain and loss in *B. castoris*

Given the differences in glycosyl hydrolase repertoires among *B. castoris* strains, I next investigated the acquisition and loss of GH families within this species based on the predictions from dbCAN2 (Figure 4.8**)**. Results suggest that during the course of evolution *B. castoris*, as well as its closest relative *B. italicum*, acquired three GH families (GH20, GH33 and GH95). These three families contain enzymes previously associated with degradation of host-derived carbohydrates in human-associated *Bifidobacterium*: lacto-*N*-biosidases, exo-sialidases and α-L-fucosidases, reported to be involved in metabolism of specific oligosaccharides, including HMOs present in maternal breast milk and intestinal glycoconjugates (*63, 158, 379, 380*). Furthermore, *B. castoris* alone, appears to have acquired GH49 family, which contains dextranases acting on dextran and pullulan (*427*). Within the *B. castoris* taxon, strain adaptation to the wild murine gut environment (as opposed to the beaver) appears to be driven by gene loss rather than gene gain events. The majority of strains isolated from hosts across all trapping sites appear to be lacking families GH51 and GH127, which predominantly contain α- and β-L-arabinofuranosidases that hydrolyse the glycosidic bond between L-arabinofuranoside side chains of hemicelluloses such as arabinoxylan, arabinogalactan, and L-arabinan (*423, 428*). The exception are 4 strains isolated from *A. agrarius* and *A. flavicollis* from site1 and site2 in Lithuania, 3 out of which appear to additionally have acquired GH46 chitosanases and lost GH20, GH33 and GH95. Overall, these observations shed light on potential evolutionary adaptations

of *B. castoris* strains to the host diet, however with only one non-mouse strain available for the analysis, the results are very speculative.

4.4.6 Identification of *eps* genes in *B. castoris*

Given the prediction of glycosyl transferase genes in the *B. castoris* glycobiome, I next examined the collection of genomes for the presence of genes potentially involved in exopolysaccharide (EPS) biosynthesis, as previous studies have indicated that EPS may support gut colonisation and stimulate host immune responses (*398*). Recently, relatively conserved genomic regions predicted to contain genes involved in EPS production have been identified in *Bifidobacterium* type strains, including members of *B. pseudolongum* phylogenetic group (gene clusters *eps*3 and *eps*4). For this search, I selected amino acid sequences of *eps* gene clusters from *B. animalis* subsp. *lactis* Bl12 (*eps*3: Bl12_1287 – Bl12_1328) and *B. pseudolongum* subsp. *globosum* LMG 11569^{T} (*eps*4: BPSG_1548 – BPSG_1565) as references (*398*).

This analysis identified homologues of several conserved *eps*-key genes in *B. castoris* genomes (Figure 4.9 & Table S4.7). All analysed strains had homologues of the *eps*3 cluster gene predicted to encode the priming glycosyl transferase (pGTF), which catalyses the first step in EPS biosynthesis. Moreover, I identified genes predicted to encode enzymes involved in rhamnose biosynthesis, showing high amino acid identity (over 80% in all cases) with glucose-1-phosphate thymidylyltransferase, dTDP-4-dehydrorhamnose 3,5-epimerase and dTDP-glucose 4,6-dehydratase from *B. animalis* subsp. *lactis* Bl12 in the collection of strains. The presence of rare carbohydrates in the EPS, including rhamnose and fucose, have been associated with additional biological properties of these polymers, including differential modulation of the immune system (*429, 430*).

Transport of the formed EPS-unit across the cytoplasmic membrane has previously been suggested to be carried out by either an ABC-type transporter or by a "flippase"-like protein (*425*). All analysed *B. castoris* genomes harboured homologues of an ABC-type transporter and I identified "flippase"-like transport homologues in strains isolated from individuals from both UK locations (Silwood and Wytham).

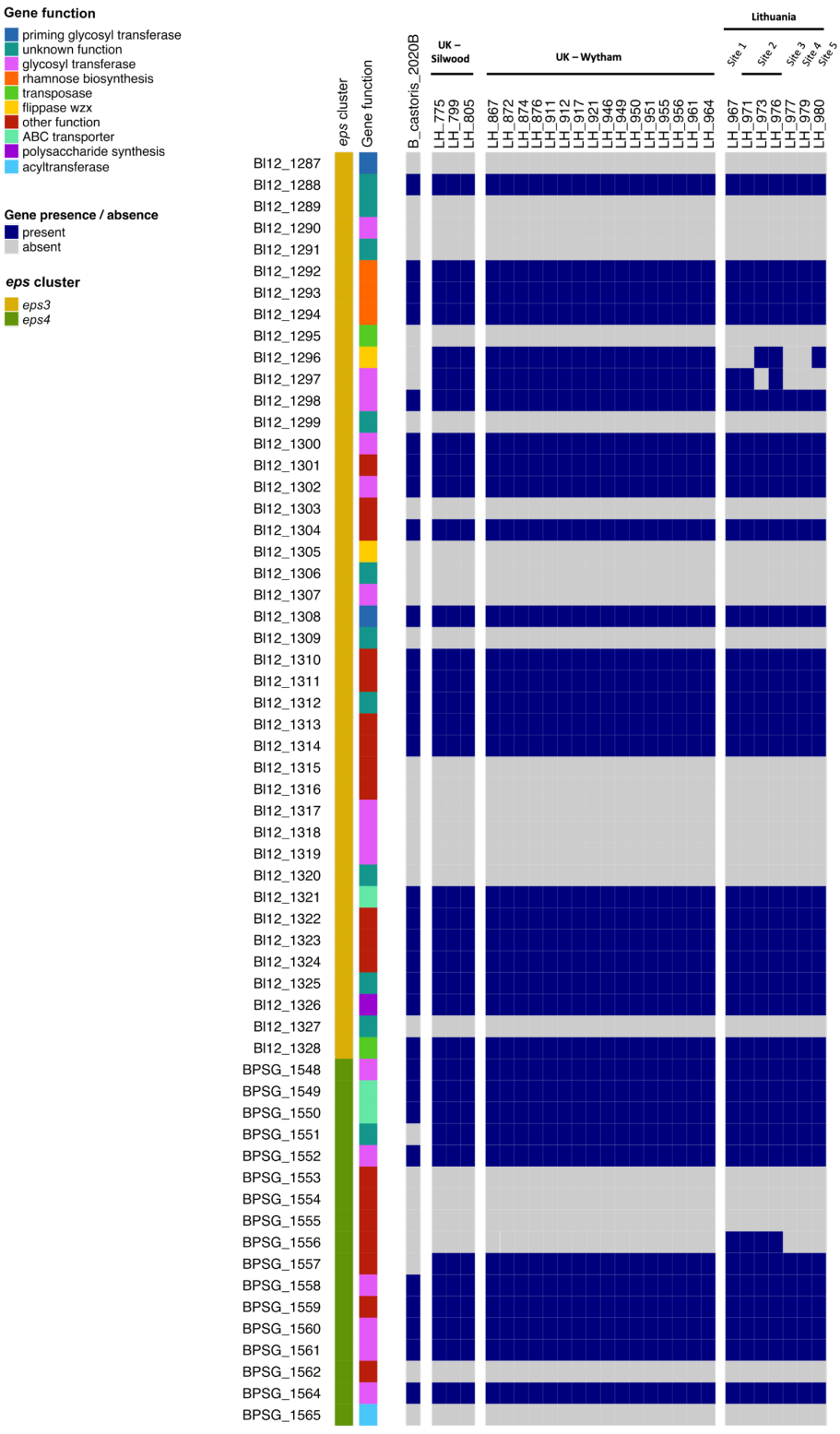


Figure 4.9 Identification of homologues of eps-key genes in B. castoris.
Sequences of B. animalis subsp. lactis Bl12 (accession number CP004053.1, eps3: Bl12_1287 – Bl12_1328) and B. pseudolongum subsp. globosum LMG 11569^{T} (accession number JGZG01000015.1, eps4: BPSG_1548 – BPSG_1565) were used as reference for BLAST+ searching.

Similarly, I identified homologues of conserved genes from *B. pseudolongum* subsp. *globosum* LMG 11569^T *eps*4 cluster in my collection of genomes, including glycosyl transferases and ABC-type transporters (Figure 4.9 & Table S4.7). These results confirm that *B. castoris* strains harbour putative *eps*-key genes and suggest potential ability for this species to produce EPS.

4.4.7 Horizontal gene transfer in *B. castoris*

Given the identification of putative *eps*-key genes in the collection of genomes, and that potential horizontal transfer of *eps* clusters in *Bifidobacterium* has previously been suggested, I next sought to determine the role horizontal gene transfer (HGT) has played in the evolution of *B. castoris* strains. For this purpose, I used the SIGI-HMM tool implemented in the software IslandViewer4. This analysis revealed that on average, 3.14±0.50 % of ORFs in *B. castoris* genomes were predicted to be horizontally acquired (range 31 to 74 genes per genome) (Figure 4.10 &Table S4.8). The highest proportion of horizontally acquired genes in each genome were those of unknown function (1.20±0.20% of ORFs and 38.46±4.36% of the predicted acquired genes on average), followed by genes involved in replication, recombination and repair (averaging 0.60±0.15% of ORFs and 19.25±4.43% of predicted acquired genes). This group, among others, encompassed genes associated with Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR). Further analysis of the HGT predictions revealed that genes involved in the cell wall/membrane/envelope biosynthesis constituted 11.35±4.35% of all predicted acquired genes per genome. This group contained genes neighbouring those identified by the BLAST+ analysis as putative *eps* genes, suggesting they might also be part of *B. castoris eps* clusters (Tables S4.7 & S4.8).

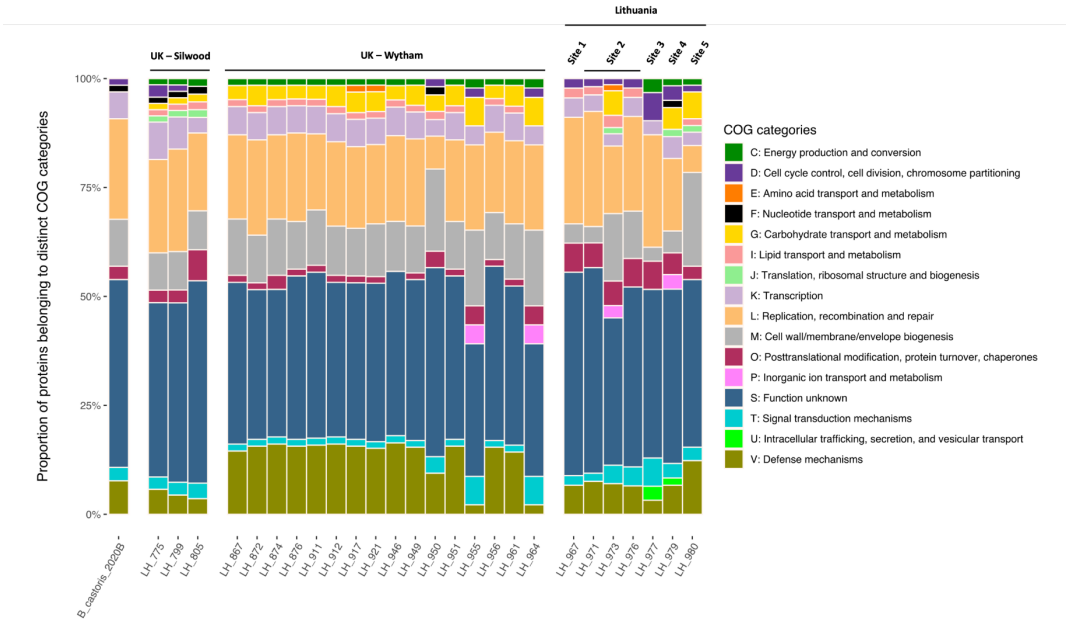


Figure 4.10 Functional classification of proteins predicted to be horizontally acquired by B. castoris strains based on COG categories.
Summary of predicted COG categories:[C] Energy production and conversion; [D] Cell cycle control, cell division, chromosome partitioning; [E] Amino acid transport and metabolism; [F] Nucleotide transport and metabolism; [G] Carbohydrate transport and metabolism; [I] Lipid transport and metabolism; [J] Translation, ribosomal structure and biogenesis; [K] Transcription; [L] Replication, recombination and repair; [M] Cell wall/membrane/envelope biogenesis; [O] Post-translational modification, protein turnover, and chaperones; [P] Inorganic ion transport and metabolism; [S] Function unknown; [T] Signal transduction mechanisms; [U] Intracellular trafficking, secretion, and vesicular transport; [V] Defence mechanisms.

Moreover, genes identified as involved in carbohydrate transport and metabolism constituted on average 3.34±2.10% of all predicted HGT genes per genome. Further analysis of the results for this group revealed that 6 strains isolated from across all locations (all strains from Silwood Park, one strain from Wytham Woods (LH_950) and two strains from Lithuania (LH_973 and LH_980)) might have acquired a GH43 family member annotated as α-L-arabinofuranosidase through an HGT event. Additionally, strains LH_955 and LH_964 from Wytham Woods, as well as strains LH_973 and LH_979 from Lithuania were predicted to horizontally acquire a GH36 family α-galactosidase (Table S4.8**)**. Overall, these results suggest that HGT may have contributed to the evolution of *B. castoris* strains and their glycobiome, however experimental validation through carbohydrate metabolism experiments

and enzyme characterisation would be essential to confirm the functional importance of these events.

4.4.8 CRISPR-Cas systems of *B. castoris*

Given the predicted acquisition of CRISPR-associated genes through HGT events, I next sought to investigate the occurrence and diversity of CRISPR-Cas systems in *B. castoris* genomes. Their presence or absence was determined using CRISPRCasFinder based on the database of known Cas protein sequences and models of CRISPR-Cas systems architecture. These systems are heritable adaptive immune mechanisms in bacteria and archaea, which originate through selection and integration of foreign nucleic acids into the genome in the form of CRISPR arrays to provide memory of infection (*431*).

Initially, I examined the occurrence of a universal core protein Cas1, required for spacer acquisition and characteristic of the CRISPR-Cas systems subtypes (*432*). Eighty five percent of *B. castoris* isolates contained *cas1* genes in their genome. This proportion was higher than previously estimated for bacteria (46%) (*433*), genus *Bifidobacterium* (up to 77%) *(434)*, and specific *Bifidobacterium* species, namely *B. longum* (38%) (*435*) and *B. pseudocatenulatum* (62%) (*436*). A total of 4 isolates, three from Silwood Park and one from Lithuania, were found to encode two *cas1* genes in different regions of their genomes, suggesting the presence of two CRISPR loci (Table 4.4). This feature has previously been reported for other strains of *Bifidobacterium,* e.g. *B. dentium* LMG 11045 (*434*), *B. longum* 1-6B, 2-2B, 44B and *B. longum* subsp. *infantis* EK3 (*435*).

Strain	Geographical origin	Type-subtype	Repeat sequence	Repeat length	No of repeats	cas1	cas3	cas9
B_castoris_2020B	Italy	No cas						
LH_775 1st locus	Silwood Park	I-E	GTGTTCCCCGCAAGTGCGGGGATGATCCC	29	51	Y		
LH_775 2nd locus	Silwood Park	II-C	CAAGTCTATCAGGAAGGGAAGAAGCTAATTTCCAGC	36	18	Y		Y
LH_799 1st locus	Silwood Park	I-E	GTGTTCCCCGCAAGTGCGGGGATGATCCC	29	51	Y		
LH_799 2nd locus	Silwood Park	II-C	CAAGTCTATCAGGAAGGGAAGAAGCTAATTTCCAGC	36	22	Y		Y
LH_805 1st locus	Silwood Park	I-E	GTGTTCCCCGCAAGTGCGGGGATGATCCC	29	37	Y	Y	
LH_805 2nd locus	Silwood Park	II-C	CAAGTCTATCAGGAAGGGAAGAAGCTAATTTCCAGC	36	13	Y		Y
LH_867	Wytham Woods	I-E	CGGATCATCCCCGCGTGTGCGGGGCAAAC	29	16	Y	Y	
LH_872	Wytham Woods	I-E	CGGATCATCCCCGCGTGTGCGGGGCAAAC	29	16	Y	Y	
LH_874	Wytham Woods	I-E	CGGATCATCCCCGCGTGTGCGGGGCAAAC	29	16	Y	Y	
LH_876	Wytham Woods	I-E	CGGATCATCCCCGCGTGTGCGGGGCAAAC	29	16	Y	Y	
LH_911	Wytham Woods	I-E	CGGATCATCCCCGCGTGTGCGGGGCAAAC	29	16	Y	Y	
LH_912	Wytham Woods	I-E	CGGATCATCCCCGCGTGTGCGGGGCAAAC	29	16	Y	Y	
LH_917	Wytham Woods	I-E	CGGATCATCCCCGCGTGTGCGGGGCAAAC	29	16	Y	Y	
LH_921	Wytham Woods	I-E	CGGATCATCCCCGCGTGTGCGGGGCAAAC	29	16	Y	Y	
LH_946	Wytham Woods	I-E	CGGATCATCCCCGCGTGTGCGGGGCAAAC	29	16	Y	Y	
LH_949	Wytham Woods	I-E	CGGATCATCCCCGCGTGTGCGGGGCAAAC	29	16	Y	Y	
LH_950	Wytham Woods	No cas						
LH_951	Wytham Woods	I-E	CGGATCATCCCCGCGTGTGCGGGGCAAAC	29	16	Y	Y	
LH_955	Wytham Woods	I-E	GTGCTCCCCGCAAGCGCGGGGGATGATCCC	29	98	Y	Y	
LH_956	Wytham Woods	I-E	CGGATCATCCCCGCGTGTGCGGGGCAAAC	29	16	Y	Y	
LH_961	Wytham Woods	I-E	CGGATCATCCCCGCGTGTGCGGGGCAAAC	29	16	Y	Y	
LH_964	Wytham Woods	I-E	GTGCTCCCCGCAAGCGCGGGGGATGATCCC	29	98	Y	Y	
LH_967	Lithuania	I-E	GTCTGCCCCGCACACGCGGGGATGATCCG	29	45	Y	Y	
LH_971 1st locus	Lithuania	I-E	GTCTGCCCCGCACACGCGGGGATGATCCG	29	47	Y	Y	
LH_971 2nd locus	Lithuania	II-C	CAAGTCTATCAGGAAGGGAAGAAGCTAATTTCCAGC	36	24	Y		Y
LH_973	Lithuania	I-E	GTCTGCCCCGCACACGCGGGGATGATCCG	29	83	Y	Y	
LH_976	Lithuania	I-E	GTCTGCCCCGCACACGCGGGGATGATCCG	29	51	Y	Y	
LH_977	Lithuania	I-E	GTGCTCCCCGCAAGCGCGGGGATGATCCC	29	61	Y	Y	
LH_979	Lithuania	No cas						
LH_980	Lithuania	No cas						

Table 4.4 CRISPR-Cas systems in Bifidobacterium castoris isolates.
Isolates harbouring complete CRISPR-Cas systems are marked in grey. Nucleotide polymorphisms between repeats in complete subtype I-E systems are marked in red.

Overall, 27 CRISPR loci were predicted in 23 *B. castoris* isolates. The designation of predicted loci into CRISPR subtypes was performed based on the system architecture and presence of signature *cas* genes - *cas3* characteristic of Type I systems and *cas9* distinctive of Type II systems. This approach resulted in the classification of 23 loci as Type I-E system and 4 loci as Type II-C system. Previous examinations of CRISPR-Cas systems in *B. longum* and *B. pseudocatenulatum* showed higher diversity of Type I subtypes in these species: in both cases subtypes I-C, I-E and I-U were identified, with subtype I-C being the most prevalent (*435, 436*). With regard to Type II systems, the presence of subtypes II-A, II-B, and II-C was previously reported for *Bifidobacterium (434)*, however neither subtype II-A nor II-B were predicted in *B. castoris*. Phylogenetic analysis based on the amino acid sequence of the Cas1 protein further supported these results and showed the divergence of the two different CRISPR subtypes grouped in two distinct branches (Figure 4.11).

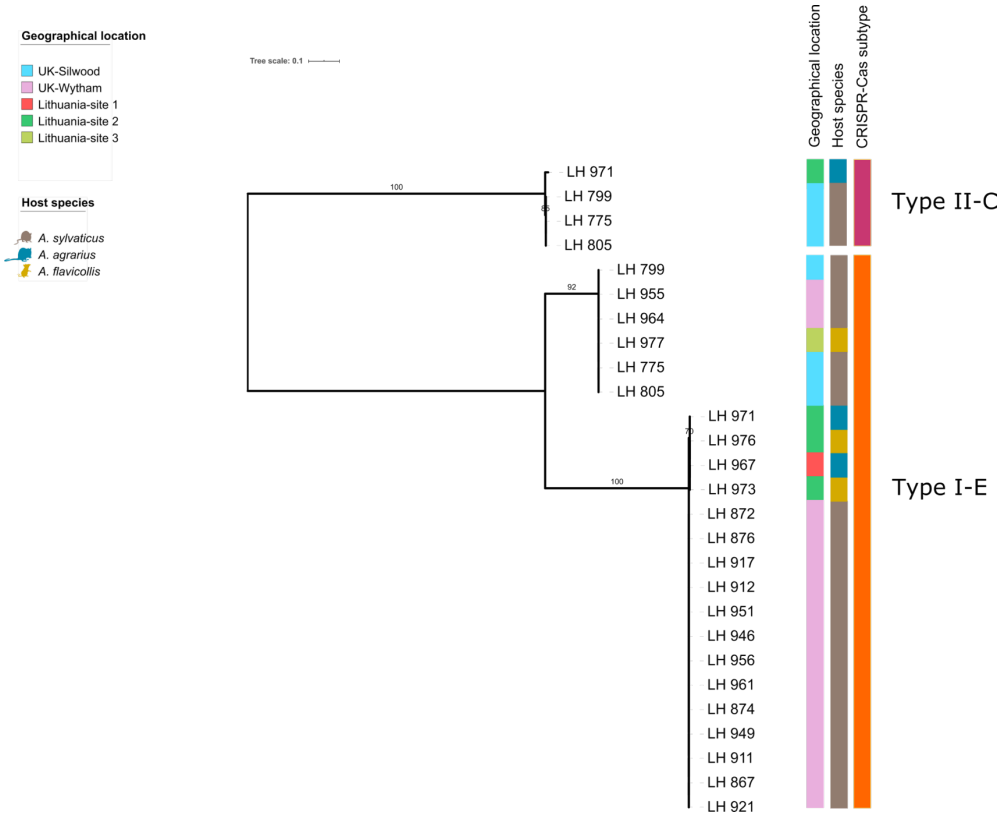


Figure 4.11 Phylogenetic tree based on the amino acid sequences of Cas1 protein in B. castoris. The tree was reconstructed using the maximum likelihood method (WAG model) with 1000 bootstrap iterations. Bootstrap values above 70 are marked on the nodes.

The 27 CRISPR loci detected in *B. castoris* isolates were next annotated with the purpose to assess their architecture (Figure 4.12). Only three strains harboured complete CRISPR-Cas systems, with strains LH_955 and LH_977 possessing subtype I-E, and strain LH_971 harbouring both subtype I-E and subtype II-C. With regard to subtype I-E, the majority of isolates (n=18) were missing *cas2* gene from the system architecture. The product of this gene, the Cas2 protein, cooperates with Cas1 in the acquisition of foreign genetic material (spacers) (*432*).

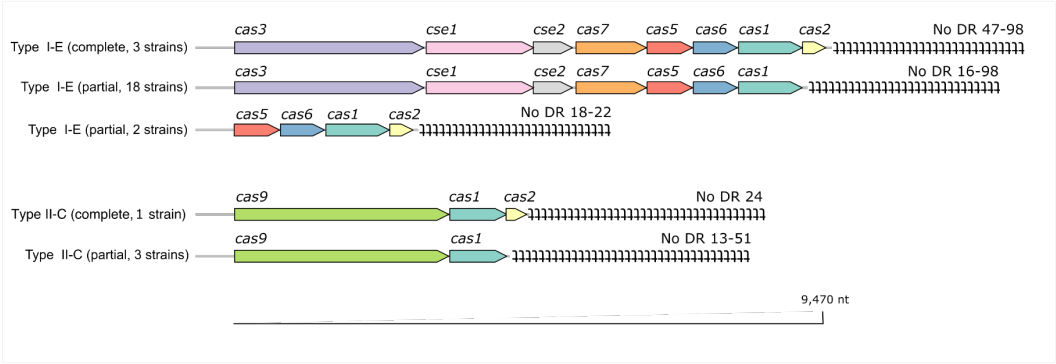


Figure 4.12 Schematic representation of CRISPR-Cas systems in B. castoris isolates. CRISPR repeats are represented as black lines on the right side of each locus (spacers are not represented).

Furthermore, 2 isolates lacked the flag subtype I-E *cas3* gene and only harboured a partial effector complex which was missing *cas7* gene. Despite the absence of the signature *cas3* gene from these CRISPR loci, they were correctly assigned to their respective system based on other distinct proteins, namely Cas1. Out of the 4 isolates predicted to additionally possess the subtype II-C system, three showed an incomplete subtype architecture missing the *cas2* gene. The presence of these incomplete CRISPR loci in *B. castoris* genomes might result either from genetic reorganization, the loss of activity toward the acquisition of the other CRISPR loci, or incomplete genome assemblies (*435*). In terms of the size of CRISPR-Cas loci, the length of complete subtype I-E systems ranged from 12 to over 15Kb due to their multi-gene architecture and high number of repeats (47-98), while the subtype II-C was shorter and encompassed lower number of accessory *cas* genes and repeats (24) (Figure 4.12). Consistent with previous reports (*435*), the length of the repeats was 29 for subtype I-E and 36 for subtype II-C. The consensus repeat sequence for complete I-E subtype in strain LH_971 showed 6 nucleotide polymorphisms in comparison to repeats from strains LH_955 and LH_977 (Table 4.4).

The CRISPR-spacer content (Figure 4.13) showed diversity across isolates consistent with the results of phylogenomic analysis (Figure 4.8) and phylogenetic investigation of the Cas1 protein (Figure 4.11) and provided additional support for previous observations on the number of strains circulating within *Apodemus* populations.

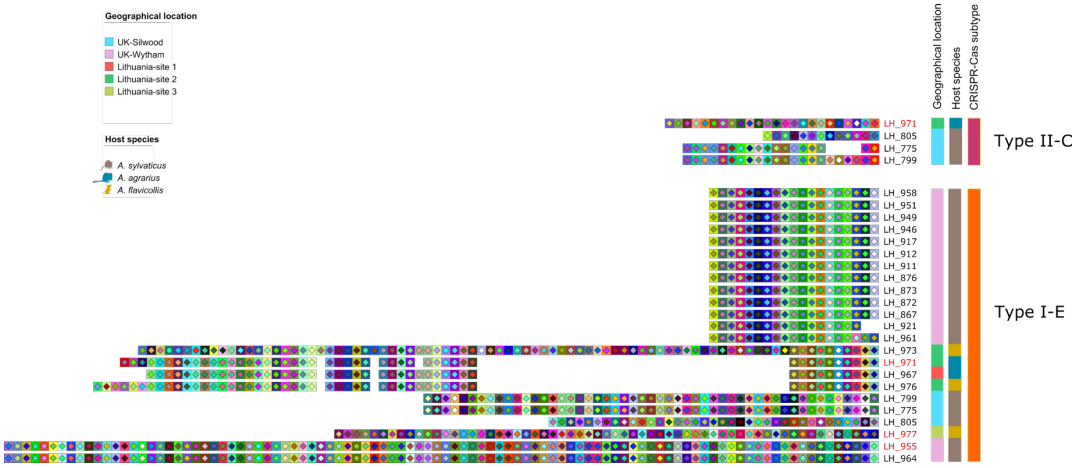


Figure 4.13 Comparison of CRISPR spacers in B. castoris.
Squares represent spacers and each unique spacer sequence is marked with unique colour and geometric figure. The last spacer acquired is represented on the left side while the first spacer is on the right side. The isolates possessing the complete CRISPR-Cas systems are marked in red.

4.4.9 Association between CPISPR-Cas and prophages in *B. castoris*

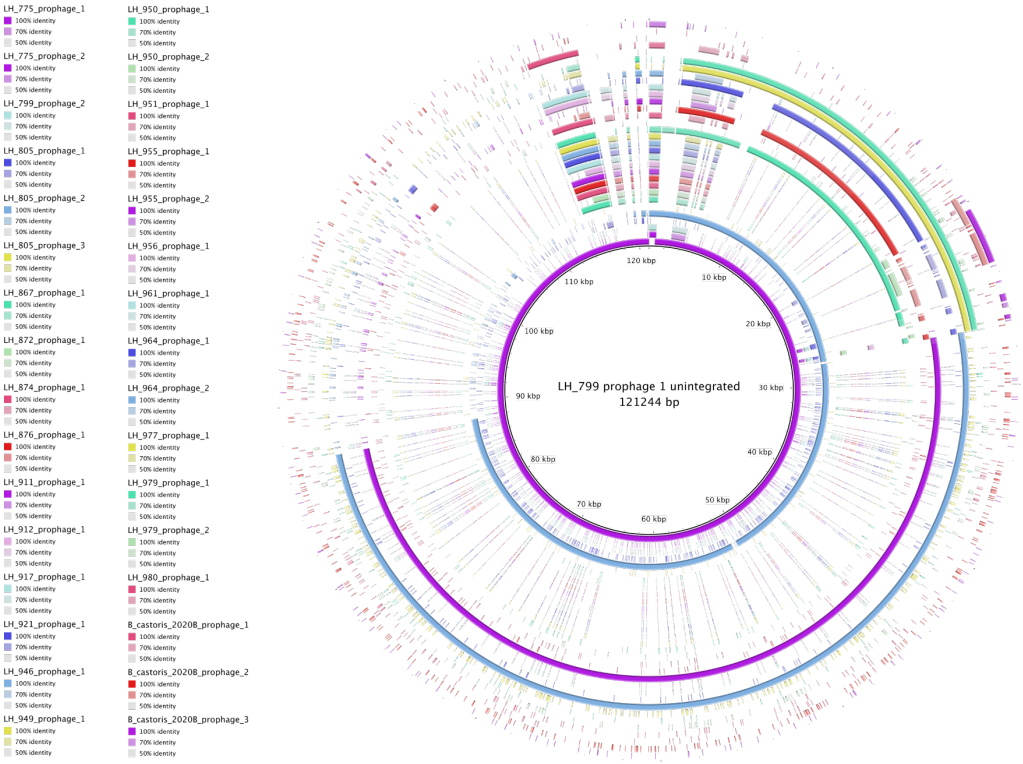
Given that spacers in CRISPR arrays originate from invading foreign elements, e.g. bacteriophages or invasive plasmids, I next sought to identify putative prophages in *B. castoris* genomes using VirSorter. This software detects viral sequences that are integrated into the bacterial chromosome (categories 4 to 6), as well as lytic and lysogenic prophages existing as an extrachromosomal plasmid (categories 1 to 3). Categories 1 and 4 contain the most confident predictions, while those belonging to categories 3 and 6 are the least certain (*436*). Here, I defined categories 1, 2, 4, and 5 as “high-confidence” and only retained predictions belonging to these categories for analysis. Additionally, I decided to refer to prophage sequences belonging to categories 1 and 2 as “unintegrated” and categories 4 and 5 as “integrated”.

The results of this analysis predicted the presence of a total of 33 viral regions in 23 out of 27 analysed isolates (Figure 4.14 & Table S4.9). Seven category 2 “unintegrated” predictions and 26 category 5 “integrated” predictions, with no predictions belonging to the highest confidence level categories (category 1

"unintegrated" and category 4 "integrated") were made in *B. castoris* isolates. Interestingly, this approach did not identify prophages in the four Lithuanian strains from site1 and site2 forming the distinct phylogenetic cluster in the *B. castoris* phylogenetic tree (LH_967, LH_971, LH_973 and LH_976) (Figure 4.7 A), which all harboured CRISPR-Cas loci. This observation suggested that the defence systems in these strains may be functional. This result may also be explained by the fact that only viral predictions classified into the "high-confidence" prediction categories in VirSorter were retained.

The size of predicted prophage sequences ranged from 16.6 to 121.3 Kb, with G+C% content between 59.42% and 67.43%. Previous findings on *Bifidobacterium*-associated prophages reported prophage sizes up to 60 Kb, with similar G+C% content (*436, 437*). Eight isolates (34.8%) had more than one prophage associated with their genome. The BLASTN comparison between predicted viral signals, with that of LH_799 prophage 1 (the longest identified) selected as "reference" (Figure 4.14A), indicated diversity among *B. castoris* prophages. To better understand their evolutionary trajectory, I constructed phylogenetic trees using whole predicted sequences, as well as the sequence for phage signature portal protein, which was identified in all prophages. In addition to the *B. castoris* prophages, whole genome sequences of two *Bifidobacterium*-associated phages Bbif-1 (GCA_002633625.1) and PMBT6 (GCA_006529735.1) were included in this analysis (Figure 4.14 B).

A)



B)

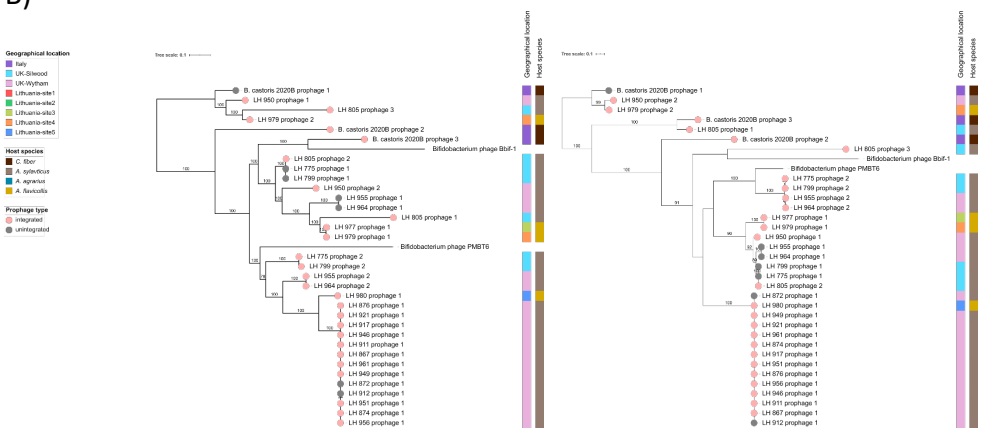


Figure 4.14 Identification of viral signal in B. castoris (A) and phylogenetic trees of B. castoris prophage elements built based on whole genome sequences (left) and portal protein (right) (B). (A) The representation of all identified sequences belonged to VirSorter categories 2 or 5, which encompass sequences divergent from references and partial sequences lacking viral hallmark genes which may include defective prophage, respectively. The map represents BLASTN comparisons of viral regions predicted for B. castoris genomes to that of an unintegrated LH_799 prophage_1. (B) Phylogenetic trees were reconstructed using the maximum likelihood method with 1000 bootstrap iterations, employing 'GTR' and 'WAG' models for whole genome sequences and portal protein, respectively. Bootstrap values above 70 are marked on the nodes.

Examination of the phylogenetic tree based on whole genome sequences indicated the presence of three main prophage clusters, while this distinction was not as clear in the topology of the tree built using the portal protein. However, there were similarities in the clustering of prophages between the two trees that were also consistent with previous findings resulting from phylogenomic analyses of the *B. castoris* taxon. Sequences of strains LH_775 and LH_779 from *A. sylvaticus* from Silwood Park, LH_977 and LH_979 from *A. flavicollis* from Lithuania, LH_955 and LH_964 from *A. sylvaticus* from Wytham Woods clustered together, indicating their close relatedness. Similarly, the clustering of the 13 prophage sequences from *B. castoris* isolates from *A. sylvaticus* from Wytham was consistent across the constructed phylogenies, supporting the previous observation on high prevalence of one particular strain in this population.

Based on the output from VirSorter, a total of 2,105 ORFs were identified in the 33 prophage sequences (Table S4.10). Out of those, the function for 1,359 (64.6%) could not be predicted (hypothetical proteins), indicating there is a vast scope for further studies. Only 203 genes (9.64%) were clearly annotated with phage-associated functions. Using previous work of Botstein (*438*) as basis, I categorised genes with predicted functions into five functional modules, namely lysogeny, DNA replication, DNA packaging, head and tail morphogenesis, and lysis (Figure 4.15). The gene encoding portal protein belonging to the DNA packaging module was the only one consistently identified in all prophages, however, this module was not very well preserved across the analysed sequences. The only other gene associated with DNA assembly was capsid protein identified in 33% of prophage sequences. Integrase-encoding gene was the second most preserved, identified in 87% of sequences, followed by genes encoding single stranded DNA binding protein and endodeoxyribonuclease, both present in 73% of *B. castoris* prophages and belonging to the DNA replication module. The lysis module was the least well represented, with lysin identified in 30% of sequences and only one prophage from *B. castoris* 2020B^{T} harbouring the gene encoding holin.

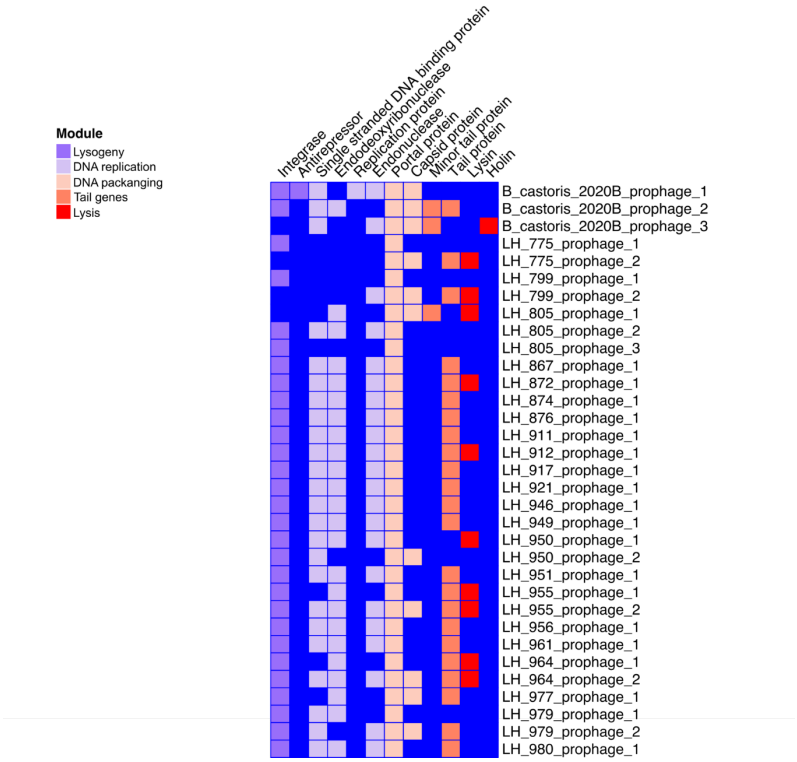


Figure 4.15 Classification of prophage genes into functional modules.
Coloured squares represent the presence of genes predicted to encode particular functional proteins in the identified prophage sequences.

Following identification of viral signals in *B. castoris*, I next sought to investigate the association between identified CRISPR-Cas systems and prophage sequences in *Bifidobacterium*. For this purpose, I screened spacers from the three *B. castoris* strains possessing the complete CRISPR-Cas systems against prophages identified in *B. castoris*, as well as the custom database of 2,030 publicly available *Bifidobacterium* genomes (downloaded from NCBI on August 6, 2020) (Figure 4.16). This investigation revealed homology between spacers from both subtype I-E and subtype II-C systems and prophage elements in *B. castoris* (Figure 4.16 A), as well as members of 7 other *Bifidobacterium* species (Figure 4.16 B). Overall, sequences from strain *B. castoris* 2020B^{T}, and in particular the integrated prophage_2, represented the most targeted prophages with matches from both subsystem I-E and II-C spacers. Further, both prophage sequences identified in the Wytham LH_950 strain were targeted by subsystem I-E spacers from strains LH_955 and LH_977, as well as subsystem II-C spacers from strain LH_971.

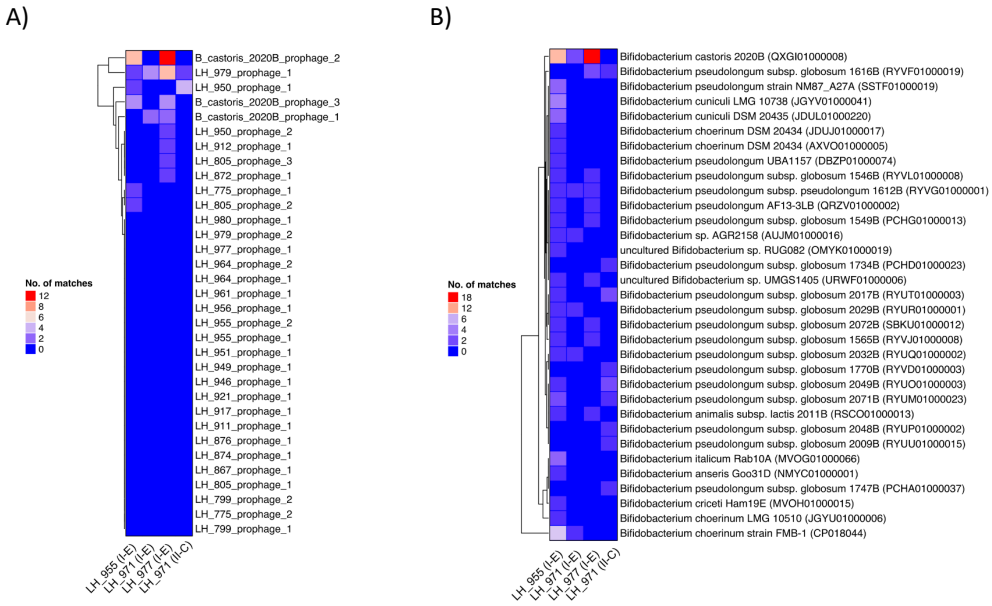


Figure 4.16 B. castoris CRISPR spacers targeting prophages in B. castoris genomes (A) and other Bifidobacterium species (B).
The vertical axis represents the prophage sequences identified in B. castoris genomes (A) and the genomes of Bifidobacterium strains that harbour prophages targeted by B. castoris CRISPR spacers (B). The horizontal axis represents CRISPR-Cas spacers from strains with complete CRISPR-Cas systems. The colour scales represents the number of targeting events.

With regard to the wider *Bifidobacterium* screen, apart from targeting prophages in type strain *B. castoris* 2020B^T, spacers from the subtype I system also targeted strains belonging to *B. choerinum*, *B. cuniculi*, *B. italicum, B. criceti*, *Bifidobacterium anseris*, *B. animalis* and *B. pseudolongum* species. Spacers from strain LH_955 had the widest range, targeting prophages in all abovementioned *Bifidobacterium* species, while spacers from LH_977 only showed homology to viral signals in *B. castoris* 2020B^T and *B. pseudolongum*. Spacers from the subtype II system only showed homology to prophage elements in *B. pseudolongum* strains. Given that *B. castoris* spacers seem to only target sequences in *Bifidobacterium* species belonging to the *B. pseudolongum* phylogenetic group (Figure 4.5) (*286*), it is reasonable to suggest that *B. castoris* strains acquired immunity against prophages capable of infecting closely related bifidobacteria.

4.5 Discussion

This work is the first to describe the results of *Bifidobacterium* screens in wild rodent populations and key aspects of the pan-genome of *Bifidobacterium castoris* species. I recovered *Bifidobacterium castoris*, *Bifidobacterium animalis* and *Bifidobacterium pseudolongum* from three species of wild mouse (*Apodemus sylvaticus*, *Apodemus flavicollis* and *Apodemus agrarius*) from two European countries, UK and Lithuania. The species *B. animalis* and *B. pseudolongum* had previously been isolated from a number of animal hosts and at the species-level are considered host generalist (*292, 293*). *B. castoris* had previously only been isolated from beavers, but here I show it also colonises mice. Some species of *Bifidobacterium* show higher host specificity than others, e.g. *Bifidobacterium tissieri* has only been associated with primates to date (*294*). While the results suggest *B. castoris* strains might be able to infect a broad range of rodents (beavers and mice so far), additional isolation efforts would be required to confirm its true host range. Interestingly, despite testing roughly equal numbers of mouse and vole faecal samples collected at trapping sites in Lithuania, I did not manage to isolate *Bifidobacterium* from voles. This could reflect either its absence in voles or presence at low abundance.

The genomic data generated in this work provided additional insight into the phylogeny and genomic features of *B. castoris*, using 26 isolates newly isolated here and the previously described type strain *B. castoris* 2020B^{T}, recovered from a captive European beaver (*Castor fiber*) in Italy. Previous research has shown strong patterns of co-phylogeny between some mammalian hosts and *Bifidobacteriaceae*. Moeller et al. (*404*) showed tight congruence between the phylogenies of *Bifidobacteriaceae* and their hominid hosts, providing support for co-diversification. In this work I tested whether a similar pattern holds in a different mammalian order, using a much higher (genome-level) resolution analysis of strains from a single *Bifidobacterium* species (*B. castoris*) and their mouse hosts. In contrast to previous work (*404*), I found no statistical congruence between the *B. castoris* and

host phylogenies. *B. castoris* strains did not show clear phylogenetic clustering by host species nor geographic region. All *B. castoris* strains identified were only found in a single mouse species, suggesting strains may be host-species specific, though more extensive sampling would be needed to test this definitively. The results are more consistent with regular host shifts among *B. castoris* strains than co-diversification, and if co-diversification did in fact occur, its signature has by now been eroded by subsequent host shifts. An interesting question is what might drive the contrasting host-*Bifidobacterium* evolutionary patterns in hominids and mice. I speculate that differences in host evolutionary history and ecology may be important here. One theoretical possibility is that speciation may have been allopatric in hominids but sympatric for *Apodemus* mice, but the literature indicates that allopatric speciation is likely in *Apodemus (439, 440)*. A more plausible explanation might be that allopatry did not persist among nascent *Apodemus* species for very long, and that since they speciated, contact among species in this genus has been more extensive than in hominids, allowing more frequent *B. castoris* transfer and host shifts. Hominid species show very strongly bounded present-day geographical ranges (*441*), whereas *Apodemus* species in Europe have very broad overlapping ranges, and different species (e.g. *A. flavicollis* and *A. agrarius* at the Lithuanian sites) can often be caught in adjacent traps. Thus, earlier and more extensive contact between *Apodemus* mouse species may have allowed more cross-species transmission and host shifts of *B. castoris* strains over evolutionary time than could have taken place among hominid species. Future work testing the generality of co-diversification across hosts groups with different speciation patterns would be highly informative to understand how the biogeographic and temporal patterns of speciation may affect evolutionary patterns in host-symbiont relationships (*412*). Another possibility is that behavioural differences have contributed to the contrasting host-*Bifidobacterium* evolutionary patterns between hominids and mice. For example, the practice of coprophagy, common to rodents, has been shown to affect the composition and abundance of the intestinal microbiota within mouse litters (*442*).

The identification of the *B. castoris* glycobiome provided insight into the strain-specific genetic repertoire predicted to be involved in carbohydrate metabolism and synthesis. The data indicated that *B. castoris* is predominantly enriched in GH families implicated in the degradation of plant-derived carbohydrates. Consistent with previous studies (*33, 38*), analysis of CAZymes identified family GH13 as predominant in all *B. castoris* genomes, constituting on average 29.98±1.88% of their GH repertoire. This finding is consistent with the largely plant-based (granivorous) diet of *Apodemus* mice, including as found at several of the sites in UK and Lithuania (*443-446*).

The predicted acquisition of family GH49 by *B. castoris* taxon is especially interesting. To my knowledge, this is the first report identifying the putative presence of enzymes belonging to this family in *Bifidobacterium*. Most known GH49 dextranases have been discovered in fungi, with only 7 listed as characterised in bacteria to date (based on CAZy database, July 2020). Overall, members of this family are not very well characterised in prokaryotes (*427*). A recent study using a recombinant dextranase from the marine bacterium *Arthrobacter oxydans* KQ11 reported positive effects of its product – an isomalto-oligosaccharide – on the growth of beneficial human-associated *Lactobacillus* and *Bifidobacterium*, and the inhibition of pathogenic *E. coli* and *S. aureus in vitro* (*427*).

Previously, several bifidobacterial arabinofuranosidases that belong to families GH43 and GH51 and act on arabinose-substituted polysaccharides have been identified in *Bifidobacterium* (*261, 447, 448*). The analysis of the GH gain loss events in *B. castoris* suggested that most strains isolated from mouse hosts lack family GH51 and GH127. The data indicated that all *B. castoris* genomes harbour 2-3 copies of GH43 arabinofuranosidases and that, interestingly, some of these copies seemed to have been acquired via an HGT event. These findings suggest there might be an evolutionary advantage for *B. castoris* in possessing GH43 arabinofuranosidases over those belonging to families GH51 and GH127, which may be linked to the composition of the host diet. However, it is difficult to speculate on biological significance of these results without the supporting experimental data. It has previously been shown that arabinofuranosidases characterised in

B. adolescentis belonging to different families display variation in substrate specificity; AbfA belonging to family GH43 removed arabinose on position C(O)2 and C(O)3 of monosubstituted xylose residues and had larger hydrolytic activity towards substrates with a low amount of arabinose substitutions, while AbfB from GH51 only hydrolysed arabinoses from the C(O)3 position of disubstituted xyloses (*261*).

Interestingly, three Lithuanian strains isolated from *A. agrarius* and *A. flavicollis* from two distinct trapping sites (site1 and site2) seemed to have acquired chitosanases (GH46). The presence of such genes may reflect nutrient availability or dietary preferences of their animal hosts, though no data on the diet of mice analysed in this study were available. Nonetheless, stomach contents analysis has previously detected fungi as a dietary item of *Apodemus flavicollis* in Lithuania at sites close to those studied here (*443*), as well as in one of the UK sampling sites. (*444*). Moreover, spores of putatively edible fungi (with macroscopic fruiting bodies) were frequently detected in the faeces *Apodemus* spp. in other parts of Lithuania (*449*), suggesting that mycophagy in *Apodemus* spp. may not be uncommon. Further information on the specific food items eaten by *Apodemus* could allow bioinformatic predictions of carbohydrate degradation properties of *B. castoris* strains, based on specific dietary components found in host diet. The presence of predicted GH46 in genomes of *B. castoris* is very interesting and has not been previously reported, to my knowledge. The potential of *Bifidobacterium* to degrade chitin-derived substrates is currently not very well understood. Previously, studies looking into functional effects of chito-oligosaccharides (COS) on gut microbiota members produced inconsistent results. Lee et al. (*450*) showed increased growth of *B. bifidum* in pure cultures supplemented with COS. Contrary to this, Vernazza et al. (*451*) did not observe any positive effects of COS on growth of human-associated bifidobacteria from faecal inocula. Furthermore, Yang et al. (*452*) reported an increase in the population of *Bifidobacterium* upon dietary supplementation of weaning pigs with COS, however a recent study using both in *vitro* fermentation and *in vivo* mouse model indicated that *Bifidobacterium* growth was significantly inhibited in mice fed with COS, leading to the conclusion that

these oligosaccharides should not be considered preferred prebiotic substrates (*453*).

These same Lithuanian strains, unlike all the remaining *B. castoris* strains, also appear to be lacking GH families containing enzymes previously associated with degradation of specific fucosylated and sialylated milk oligosaccharides and intestinal glycoconjugates in *Bifidobacterium* isolated from human hosts (GH20, GH33, GH95). The structure and composition of the milk oligosaccharides differ greatly between mammalian species. According to Prieto et al. (*454*) human milk contains the most complex mixture of reducing oligosaccharides, many of which include fucose and determinants for human blood groups, e.g. the ABO and Lewis system. Compared to human breast milk, composition of oligosaccharides in mouse milk is mainly limited to sialyllactoses with minuscule amounts of fucosylated lactose (3'-fucosyllactose) (*454*). The fact that I identified enzymes belonging to families GH20, GH33 and GH95 in *B. castoris* strains indicate that members of this species may have the ability to metabolise specific host-derived oligosaccharides, including milk oligosaccharides, however experimental evaluation of substrate specificity of these enzymes is required.

The identification of genes predicted to encode GTs in genomes of *B. castoris* prompted questions about potential ability of members of this species to produce EPS. Recent genomic studies have described high levels of inter-species variation with respect to the number, function and organisation of genes in *Bifidobacterium eps* clusters (*398, 425*). However, a set of conserved *eps*-key genes has been proposed as universal markers, including genes predicted to encode the priming glycosyl transferase (pGTF), other glycosyl transferases (GTs), transporter enzymes (either "flippases" or ABC transporters) and various carbohydrate precursor biosynthesis or modification enzymes (*425, 455*). The results of the BLAST+ search for the homologues of these *eps*-key genes previously identified in type strains most closely related to *B. castoris* 2020B^{T} revealed their presence in all *B. castoris* strains, suggesting this species may be able to synthesize EPS. Furthermore, the analysis of predicted HGT events identified additional genes of unknown function neighbouring the *eps*-key genes that may be part of a distinct *B. castoris eps* cluster.

These findings support previous suggestions on a possible role of HGT in acquisition of complete or partial *eps* clusters by *Bifidobacterium* (*398*), however additional studies are required to assess the functionality of the putative EPS biosynthesis machinery in *B. castoris* species. Previous analyses of the genomes of *Bifidobacterium* type strains reported that each taxon, except for *B. bifidum*, possess at least one *eps* cluster (*398*). In line with these observations, I identified homologues of protein members of clusters *eps3* and *eps4* in *B. castoris*, including enzymes involved in rhamnose biosynthesis in cluster *eps3*. These results may suggest that EPS polymers encoded by different clusters may exert different effects on the gut environment. Previous studies showed that rhamnose-rich EPS from *Lactobacillus paracasei* DG displayed particular immunostimulatory properties by enhancing the gene expression of the proinflammatory cytokines in the human monocytes (*430*), while polymers mainly composed of glucose and galactose have been associated with modulation of gut microbial structure and metabolism by acting as dietary components (*456, 457*).

CRISPR-Cas systems provide bacteria with an evolutionary advantage, acting as defence mechanisms against invasion from either prophages or other types of foreign DNA. The analysis of CRISPR-Cas systems in *B. castoris* revealed the presence of two different subtypes belonging to Type I and Type II systems. Previous investigations of CRISPR-Cas systems in human-associated *B. longum* and *B. pseudocatenulatum* reported higher diversity of subtypes, with Type I systems represented by I-C, I-U and I-E subtypes in both species and Type II systems represented by subtypes II-A and II-C in *B. pseudocatenulatum* and *B. longum*, respectively (*435, 436*). The majority of *B. castoris* isolates were found to harbour subtype I-E, while subtype II-C was detected in 4 strains, with only one strain possessing a complete locus. Among CRISPR-Cas types, Type I-E was previously reported as the most prevalent in genus *Bifidobacterium* based on the analysis of 48 type strains (*434*). Type II systems, however, have been found to be the least common in nature (*458*), and in genus *Bifidobacterium* (*434*). In line with these findings, the presence of Type II systems in *B. castoris* is uncommon and seems to be a strain-specific characteristic rather than a general feature.

Given that CRISPR arrays are hypervariable, yet conserved genomic regions, their use as an additional tool for strain genotyping and evolutionary studies has previously been postulated (*434, 435*). Indeed, the analysis of CRISPR spacers showed diversity across *B. castoris* that was in line with the results of phylogenomic analysis and further supported identification of strains circulating within *Apodemus* populations. Further, spacer diversity indicated that these elements seemed to have originated from various sources, suggesting that the *Apodemus* gut is a phage-rich environment. The spacers from strains harbouring the complete CRISPR-Cas systems were predicted to target prophages present not only in *B. castoris*, but also in 7 other *Bifidobacterium* species. Interestingly, all targeted strains belong to *B. pseudolongum* phylogenetic group, which indicates they may share the same ecological niche in their respective hosts that is also rich in phages.

Efforts to identify prophages in *B. castoris* resulted in the detection of viral signals in 80% of analysed isolates. All identified prophage sequences were categorised as either divergent from viral references or partial and potentially lacking viral key genes. Indeed, sequence analysis indicated the presence of some of the phage-associated signature genes in predicted prophages, including portal protein, integrase and genes involved in DNA replication. Lysis was the least preserved functional module in identified prophages, suggesting they may not be able to enter into the lytic cycle (*459*). However, many of the well-established hallmark genes, e.g. those encoding head protein, terminase or tape measure protein, could not be identified based on available annotation, suggesting that *B. castoris* prophages may be defective. Prophage degeneration has previously been shown to be a common process under purifying selection (*460-462*). Previous studies on *E. coli* indicated that despite undergoing inactivation followed by much slower degradation after being incorporated into bacterial genomes, defective prophages can still have adaptive functions, such as production of non-infective particles (e.g. bacteriocins) (*462*). Further bioinformatic analyses incorporating data from several software for better gene annotation combined with experimental approaches would be required to assess functional properties of prophages in *B. castoris*.

4.6 Future work

Although it is now well established that members of *Bifidobacterium* have a wide host range, studies focusing on members of this genus in animals remain lacking. This work has laid the foundations for further examinations of *Bifidobacterium* communities in wild host populations. Additional screening of other wild host populations across different geographical locations for the presence of *Bifidobacterium* would help assess the distribution and prevalence of members of this genus, and perhaps lead to the discovery of new species. With regard to *B. castoris* specifically, such efforts would help establish whether this species is restricted to rodent hosts. This could be achieved in the laboratory setting by further isolation experiments. In addition, generation of metagenomic data using shotgun metagenomic sequencing and the use of genome reconstruction methods would complement laboratory approach. These data would also allow examination of prokaryotic and viral communities in wild host populations.

The prediction of glycosyl hydrolase families GH46 and GH49 in *B. castoris* – previously unreported for members of *Bifidobacterium* – presents potential for further investigation. Carbohydrate metabolism experiments using chitosan and dextran as carbon sources would complement the results of bioinformatic analysis, and if the functionality is confirmed, further transcriptomic and proteomic analysis would allow to determine genes and gene clusters involved in the utilisation of specific carbohydrates. Bacterial co-culture experiments would allow to assess the effect of the degradation products on other members of the gut bacterial community. Additionally, characterisation of catalytic activity of recombinant glycosyl hydrolases could reveal potential industrial applications for these enzymes (e.g. in the prebiotic industry).

Similarly, experimental procedures would confirm the functionality of predicted *eps* clusters. Phenotypic characterisation of strains producing surface-associated EPS could be performed using ruthenium red-milk medium. Furthermore, transcriptomics of strains cultured in mono- or multi-association in different types of media (e.g. faecal-based vs. non-faecal based) would allow to assess how

co-culture and medium type stimulate the transcription of genes involved in EPS production. Additionally, the functional effects of EPS-producing strains, as well as extracted EPS, on the cells of the immune system and the intestinal epithelium could be investigated experimentally, e.g. measurement of cytokine responses in stimulated immune cells or adhesion studies using fluorescent probes.

Potentially, Type II CRISPR-Cas systems can be exploited to create new tools for genome editing and engineering. The characterisation of the complete Type II CRISPR-Cas system in this study would be complemented with the bioinformatic analysis of protospacer adjacent motifs (PAMs) essential for Cas9 target recognition. Experimental approaches using transcriptomics and associated bioinformatic analysis could be used to investigate whether the identified CRISPR-Cas systems are actively transcribed and protect against invasive DNA, e.g. prophages.

The investigation of prophages in *B. castoris* genomes would be further expanded by integration of data from other available software to improve the identification of prophage-associated genes. These data could then be used to assess whether experimental prophage induction using Mitomycin C is possible, and if successful, electron microscopy could be used to confirm the structure of isolated phages. Whole genome sequencing of phage DNA would significantly expand the currently very limited data on *Bifidobacterium*-associated phages.

Chapter 5

Genomic signatures of animal-derived *Bifidobacterium* are associated with their isolation sources.

Animal faecal samples were obtained through collaboration with Ms Sarah Goatcher at Banham Zoo and Africa Alive (Norfolk, UK) in years 2014 - 2017.

Between June 2014 and October 2016, bacterial isolations and associated DNA extractions and processing for the purpose of initial identification based on 16S rRNA gene and whole genome sequencing of isolates initially identified as *Bifidobacterium* sp. were carried out by Ms Jennifer Ketskemety and Ms Charlotte Leclaire. I continued this work as part of my PhD.

DNA library preparation for WGS was done by sequencing team at the Wellcome Sanger Institute (Hinxton, UK).

I performed all the genomic and phylogenomic analyses.

5.1 Introduction

The widespread distribution of *Bifidobacterium* in the animal kingdom is well documented. However, compared to human-derived strains, animal-associated isolates are significantly underrepresented in databases. In this study, we isolated *Bifidobacterium* from faecal samples of captive animals obtained in collaboration with the Banham Zoo and Africa Alive (Norfolk, UK), and performed whole genome sequencing and bioinformatics analysis. Experimental efforts resulted in the recovery of a substantial number of *Bifidobacterium* isolates from a wide range of hosts, including a subset of strains identified to belong to putative novel species. Bioinformatic analyses indicated an "open" pan-genome of the genus, suggesting more novel genes to be discovered. In addition, *Bifidobacterium* isolated from different hosts and environments displayed differences in genomic signatures such as genome size and particular traits related to metabolism of carbohydrates.

5.2 Background

In recent years, the importance of next-generation sequencing for studying bacterial diversity has increased significantly. This approach allows for next stage genomic, phylogenomic and evolutionary analyses that facilitate taxonomic classification of bacterial isolates and provide insights into the mechanisms underlying colonisation and establishment of a particular organism within a more complex microbial community. However, the selection of organisms for sequencing has largely been guided by their public health or applied industrial relevance (*463*). In case of members of the genus *Bifidobacterium*, human-associated strains, in particular, have received much attention in recent years due to their probiotic properties.

Since the publication of the first genomic sequence of *Bifidobacterium* in 2002 (*B. longum* subsp. *longum* NCC2705) (*44*), the number of available bifidobacterial genomes has been growing rapidly. However, although over 80 *Bifidobacterium* species have currently been identified, sufficient genomic information for the majority of representative taxa is lacking. Overall, human-derived species are overrepresented in databases, in particular those associated with infant hosts and thought to exhibit health-promoting effects (e.g. *B. longum*), while those associated with animals are in the minority despite recent additions of newly characterised *Bifidobacterium* species and strains (*292, 293, 413, 464, 465*). As a consequence, broad information on the intra-species genomic variation is mainly available for those overrepresented *Bifidobacterium* species (*36, 61, 466*). In addition, studies that correlate *Bifidobacterium* genomes to the host or environment from which the sequenced isolates were recovered, and assess whether these genomes display particular environmentally relevant features or signatures of host specialisation, are limited (*31, 463*).

A more robust representation of animal-associated *Bifidobacterium* strains and species would improve the analytical power of genomic methods, and facilitate the understanding of specific genomic patterns across different species, as well as particular habitat-specific drivers behind genome evolution (*463*). In addition, this

extra genomic information would help assess the true extent of diversity within the genus (*463*). For this purpose, as many *Bifidobacterium* genomes as possible should be sequenced and annotated, across a wide host range.

To date, a number of studies have sought to determine *Bifidobacterium* diversity (*25, 27-31*). However, the majority of these studies have either focused on a subset of human-associated isolates or examined single type strains as representatives of species, and in consequence included a limited number of the isolates from different habitats. Interestingly, all studies suggest an "open" *Bifidobacterium* pan-genome (*25, 27-31*). However, since the number of genomes included in the pangenomic approach may affect the accuracy of prediction (*467*), it is important to re-assess the pan-genome of the genus *Bifidobacterium*, and explore their intra- and inter-species diversity using a higher number of available genome sequences.

This work sought to provide new insights into the diversity of animal-associated *Bifidobacterium*. I undertook isolation of new species and strains of *Bifidobacterium* from a wide range of hosts, covering mammals, birds, reptiles and insects. This effort resulted in the recovery of over 100 unduplicated isolates, including strains representative of 19 putative novel *Bifidobacterium* species. Preliminary analysis of the associated genomic data indicated an open nature of the pan-genome of the genus and suggested the presence of significant differences in genomic signatures between strains associated with particular animal host groups and environments.

5.3 Hypothesis and aims

The diversity of *Bifidobacterium* in animal hosts is not fully explored. Experimental and computational methods (bacterial isolation and genomic analysis, respectively) contribute to the discovery of true diversity within the genus. Whole genome sequences of animal-associated isolates display particular traits related to their isolation source and functional capabilities, e.g. degradation of carbohydrates.

Aims:

1) Isolate *Bifidobacterium* from faecal samples of captive animals and sequence their genomes
2) Define and create the genomic dataset composed of unduplicated newly sequenced genomes as well as publicly available sequences, with particular focus on animal-associated isolates
3) Determine genomic features of isolates in the dataset, with particular focus on traits directly and indirectly linked to carbohydrate metabolism
4) Further explore genomes of selected strains for other interesting properties e.g. defence systems

5.4 Results

5.4.1 Notes on the isolation of bifidobacterial species from animal gut microbiota samples

At the start of my PhD, the zoo animal microbiome project was already in progress, with 240 faecal samples from 116 species of mammals, birds, reptiles and insects processed for bacterial isolation by former lab members between June 2014 and October 2016. The isolation protocol for the culture of bacterial isolates from faecal samples involved the use of MRS agar supplemented with mupirocin (50mg/l) and cysteine (50mg/l) as selective medium (MRSCM). This medium, officially termed bifidobacterial selective medium, was developed by Leuschner et al. (*468*) for the enumeration of bifidobacterial species from animal feed. It has been shown to inhibit growth of *Lactobacillus*, *Streptococcus* and *Bacillus* strains, and reduce the colony size of *Pediococcus*, *Enterococcus* and *Propionibacterium*, allowing these species to be visually distinguished from *Bifidobacterium*.

However, previous experiments conducted by my colleagues, in which MRSCM was used for the isolation of *Bifidobacterium* from animal faecal samples, produced a number of "false positive" results. Several bacterial species have previously been observed by lab members to have similar morphology to that of bifidobacteria, including clostridia at early growth phases and enterococci, when grown on this selective medium. Consistent with this observation, I identified 44 isolates out of 66 available for DNA extraction at the start of my PhD as bacteria other than bifidobacteria (66%), out of which 27 belonged to the class Clostridia, and 8 were enterococci (Table S5.1). Thus, I decided to introduce an additional selective agent, sodium iodoacetate, into the isolation procedure, with the aim of increasing the selectivity for bifidobacteria.

Sodium iodoacetate contains one iodine atom attached to its methyl group and is a potent inhibitor of the glycolytic enzyme glyceraldehyde-3-phosphate dehydrogenase (GAPDH). Previous studies in lactic acid bacteria showed that inhibition of GAPDH activity with iodoacetate caused a shift towards homolactic fermentation in a mixed-acid-producing strains (*298, 299*). *Bifidobacterium*, being

heterolactic, do not have an obligate requirement for GAPDH activity in their carbohydrate metabolism pathway (*37*). Thus, sodium iodoacetate has been used in a number of selective media formulations, developed for isolation and enumeration of bifidobacterial species (*469, 470*).

To test if different media types with/without sodium iodoacetate could provide increased selectivity for *Bifidobacterium*, I cultured three animal faecal samples on BHI, MRS and RCA supplemented with, and without this selective agent (Figure 5.1). Sample Z200 was recovered from the lab isolate archive and had previously been confirmed to contain bifidobacteria, whereas samples Z241a and Z243 were more recently obtained from the Banham Zoo and were due to be processed.

BHI and RCM are very rich media, and visibly supported growth of a variety of bacterial species, regardless of whether sodium iodoacetate was added to the medium or not (Figure 5.1). Most of the BHI- and RCA-based plates contained bacterial species with distinguishably different colony morphologies, including ones with colony morphology consistent with that of *Bifidobacterium*. Similar observations were made for MRSCM medium. MRSCM+I medium, however, was observed to either support growth of bacteria with colony morphology consistent with that of bifidobacteria, but significantly smaller in size (Z200 and Z241a), or did not support bacterial growth at all (Z243).

Although media containing sodium iodoacetate have widely been used for isolation of bifidobacteria, a number of research groups have reported selective bifidobacterial growth inhibition linked to iodoacetate concentration. Silvi et al. (*471*) compared three selective bifidobacterial media, and found that BIM-25 based on RCM and containing sodium iodoacetate at 25mg/l inhibited growth of *B. longum*, and failed to recover bifidobacterial species from human and rat faecal samples. Similarly, Munoa and Pares (*469*) found that this medium inhibited *B. adolescentis* strains. Other groups reported that bifidobacterial growth was restored when the concentration of sodium iodoacetate was decreased (*472*).

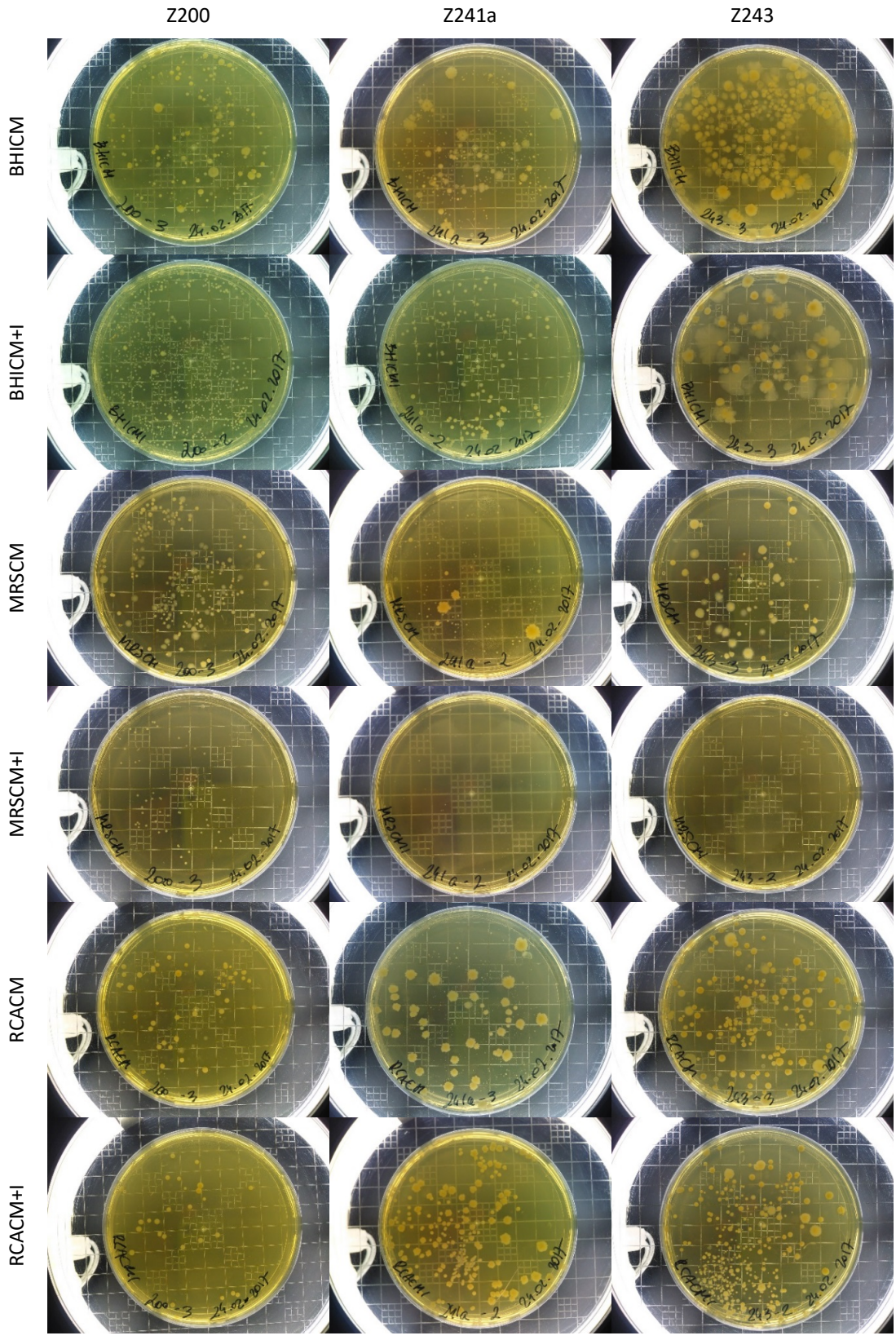


Figure 5.1 Growth of bacteria from three animal faecal samples (Z200, Z241a and Z243) on RCM agar, MRS agar and BHI agar with the addition of cysteine (C) 50mg/l and mupirocin (M) 50mg/l (labelled RCACM, MRSCM and BHICM, respectively) and additionally supplemented with sodium iodoacetate (+I) at 25mg/ (labelled RCACM+I, MRCSM+I and BHICM+I).

Thus, I decided to screen the *Bifidobacterium* library available in the lab for growth inhibition on MRSCM and BHICM with addition of sodium iodoacetate at 3 different concentrations (25mg/l, 15mg/l and 7.5mg/l). These media were selected based on the isolation results obtained in the lab prior to the start of my PhD, where distinct bifidobacterial species were recovered from both human and animal samples using MRSCM (at least 15 distinct bifidobacterial species), as well as on the results of my own experiments with different media types (Figure 5.2).

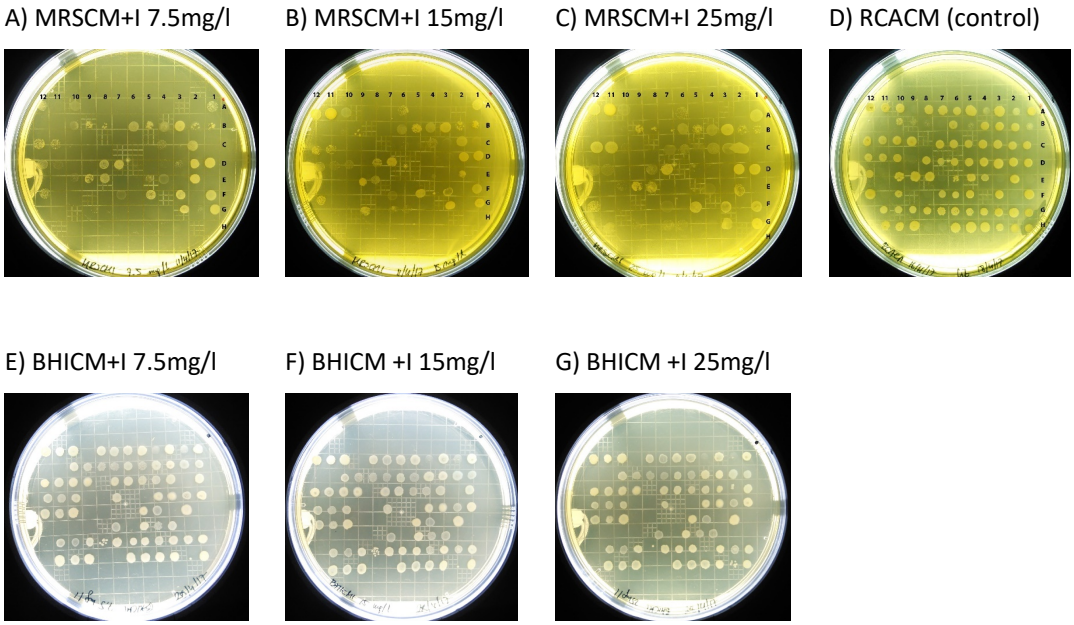


Figure 5.2 Growth of different species and strains of Bifidobacterium on MRS agar (top panel, A-C) and BHI agar (bottom panel, E-G) supplemented with cysteine 50mg/l (C), mupirocin 50mg/l (M) and sodium iodoacetate (+I) at three different concentrations: 7.5mg/l, 15mg/l and 25mg/l (marked on the figure, respectively). RCM agar (RCA) (top panel, D) with the addition of cysteine (C) 50mg/ and mupirocin (M) 50mg/l was used as control.

From the results shown in Figure 5.2 A-C, the combination of MRS agar with mupirocin (50mg/l), cysteine (50mg/l) and sodium iodoacetate, regardless of the concentration of the latter, seemed to selectively inhibit the growth of a significant number of *Bifidobacterium* strains isolated both from human and animal hosts (compared to control, Figure 5.2 D). Bifidobacterial capacity to grow on MRS with sodium iodoacetate appears to be largely strain specific. Interestingly, none of the

human-associated isolates identified with BLAST as *B. pseudocatenulatum* grew on any of the plates containing MRSCM+I (e.g. row H, wells 3-6 and 10-12).

Similarly, strain-specific growth was observed when the library was grown on the medium composed of BHI agar supplemented with mupirocin (50mg/l), cysteine (50mg/l) and different concentrations of sodium iodoacetate, however the overall inhibitory effects of these medium formulations on bacterial growth were weaker (compared to MRSCM+I) (Figure 5.2 E-G). The isolates showed identical growth profiles regardless of the concentration of sodium iodoacetate. Based on the results of this experiment I concluded that BHI- but not MRS-based medium supplemented with sodium iodoacetate could potentially be used in future isolation experiments that pose particular difficulties in terms of *Bifidobacterium* recovery, with the caveat that the diversity of the recovered strains and species may be affected.

5.4.2 Defining the study population

Prior to the start of my PhD, 60 isolates recovered from 20 animal species were identified as *Bifidobacterium* spp. based on partial 16S rRNA gene and subsequently subjected to whole genome sequencing. A further 66 isolates were available for DNA extraction and initial identification. I carried out these experiments shortly after starting my PhD. Preliminary analysis of partial 16S rRNA gene sequences using the BLAST algorithm identified 18 out of the 66 sequences (27%, from 13 animal species) as bifidobacteria and one insect-derived sequence inconclusively as either *Neoscardovia arbecensis* or *B. breve*. *N. arbecensis* has been proposed as the only representative of new genus *Neoscardovia* in the family *Bifidobacteriaceae* in 2012 based on the analysis of 16S rRNA gene and *hsp60(groL)* gene sequences, with a maximum identity of 94% to its closest relatives, including bifidobacterial species, namely *B. indicum* JCM 1302^T and *B. breve* DSM 20213^T (*473*). Due to the close relationship between the two species, I decided to treat inconclusive BLAST results as potential *Bifidobacterium* isolates to be validated by whole genome sequencing.

Between October 2016 and November 2017, I received and processed a further 109 faecal samples from 98 animal species. These efforts resulted in the preliminary

identification of an additional 88 isolates from 17 animal species, putatively identified as *Bifidobacterium* strains. In total, 167 samples from 51 animal hosts were subjected to whole genome sequencing within this project. Additionally, 33 human-associated isolates and one isolate from a probiotic formulation were added to this collection, bringing the total number of sequenced samples to 201.

Following sequencing, contamination check and the initial ANI analysis, I excluded 81 samples, as either contaminated or duplicate sequences (ANI > 99.99%, isolated from the same faecal sample), with 120 genomes from 40 host species retained for further analysis (Table S5.2). The assembled draft genome sizes for newly sequenced isolates ranged from 1.85 Mb to 3.27 Mb, with G+C% content between 54.83% and 65.88%. These values are in line with previous reports for members of genus *Bifidobacterium* (*27*).

Additionally, 281 publicly available *Bifidobacterium* genome assemblies were downloaded from NCBI and included in the dataset, together with the *B. castoris* data generated for the wild mammal study (Chapter 4). The vast majority of genomes available in the NCBI database represent human isolates, and in many cases there is very limited metadata available (e.g. information on host age and environmental background of the sample (e.g. faeces, milk, oral cavity, urogenital, etc.)). Since the primary focus of this project was on animal-associated *Bifidobacterium*, I decided to select 5 representative publicly available sequences from each human-associated *Bifidobacterium* species and subspecies, and tried to include both adult and infant isolates where possible. The final dataset encompassed 433 genomes representing isolates obtained from a wide range of hosts, both animal (n=350) and human (n=68), as well as loosely defined "environmental" sources (anaerobic digester, dairy, probiotic formulation and sewage) (n=13) (Figure 5.3). The majority of genome assemblies included in the dataset were of very high quality (91.5%, less than 50 contigs).

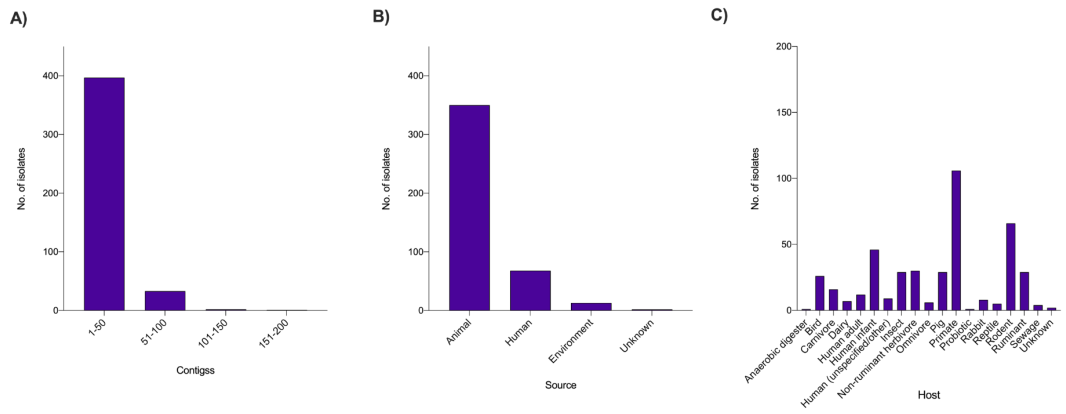


Figure 5.3 Statistics of genomes included in the analysis.
(A) Number of contigs in the assemblies. (B) Isolate categories according to animal/human/environmental sources. (C) Isolate categories according to host/environment groups.

Animal- and human-associated *Bifidobacterium* genomes were assigned to arbitrary host categories at the intersection of taxonomic classification and diet. Non-human mammalian isolates, which constituted the majority of the dataset, were split into groups based on the type of diet i.e. omnivore, carnivore, herbivore (ruminant and non-ruminant). Taxonomic grouping was employed where a considerable number of isolates representing animals at a specific taxonomic level were available (e.g. primate n=106, rodent n=66, pig n=29, rabbit n=8). Human-associated isolates for which the corresponding metadata were available were grouped into adult and infant categories (Figure 5.3).

5.4.3 Preliminary analysis of potential new *Bifidobacterium* species

Comparison of ANI values between 120 isolates and 87 *Bifidobacterium* type strains allowed taxonomic assignment of 76 genomes (63.33%) to 31 *Bifidobacterium* species and subspecies (ANI > 95%) (Table S5.3). The predominant species and subspecies identified through this analysis included animal-associated *Bifidobacterium pseudolongum* subsp. *globosum* (7 isolates, 9.21% of assigned genomes), *Bifidobacterium reuteri* (also 7 isolates, 9.21% of assigned genomes) and *Bifidobacterium animalis* subsp. *animalis* (5 isolates, 6.58% of assigned genomes). Surprisingly, genomes of 44 isolates (36.67% of sequenced samples) showed ANI

values below 95% to the known *Bifidobacterium* type strains, suggesting they may represent putative novel species. These isolates were recovered from 21 animal hosts, with those obtained from primates constituting the highest proportion (52.3% of putative novel isolates). Other hosts harbouring potential novel *Bifidobacterium* species included a South American species of rodent – Azara's agouti (18.2% of isolates), various species of birds (13.6%), a cockroach (9.1%) and two species of tortoises (6.8%). Interestingly, a number of isolates within this group, both from the same and different hosts, displayed ANI values above 95% between each other (Figure 5.4). Further analysis of these data revealed that the 44 isolates appeared to belong to 19 putative novel *Bifidobacterium* species.

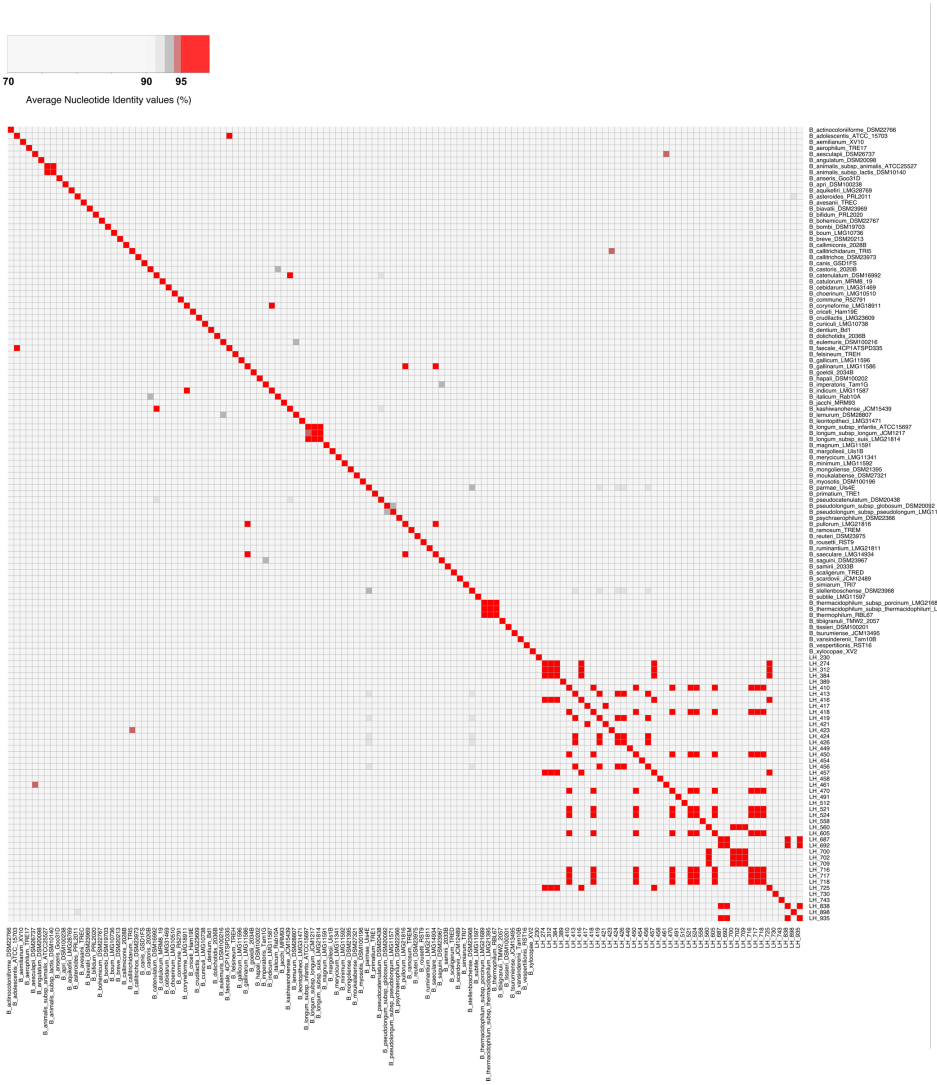


Figure 5.4 ANI analysis between the type strains of the recognised 87 species of Bifidobacterium and the isolates predicted to belong to putative novel Bifidobacterium species. The diagram shows values between 70-100%, values above 95% are marked in red.

Based on these findings, I sought to conduct a preliminary analysis of these putative novel species using additional *in silico* methods. For this purpose, I predicted their 16S rRNA gene sequences and performed a screen against the SILVA 16S rRNA gene sequences database (v.138, 16 December 2019) (Table 5.1).

Isolate	Host	Group	Host ID	ANI value type strain (>95%)	ANI type strain	Same species (ANI > 95%)	16S rRNA species (SILVA database)	16S rRNA identity (%)	16S rRNA coverage (%)
LH 230	Cotton top tamarin	Primate		n/a	Potential novel	1	FNSE01000001.1499458.1500992_Bifidobacterium_dentium_JCM_1195	94.792	100
LH 274	Cotton top tamarin	Primate		n/a	Potential novel	2	FTRS01000105.3844.5341_Bifidobacterium_longum	98.675	100
LH 312	Emporer tamarin	Primate		n/a	Potential novel	2	FNRW01000001.379172.380700_Bifidobacterium_longum	98.317	100
LH 384	Emporer tamarin	Primate		n/a	Potential novel	2	FNRW01000001.379172.380700_Bifidobacterium_longum	98.317	100
LH 416	Cotton top tamarin	Primate		n/a	Potential novel	2	FNRW01000001.379172.380700_Bifidobacterium_longum	97.253	100
LH 457	Emporer tamarin	Primate		n/a	Potential novel	2	JRWN01000007.321.1855_Bifidobacterium_longum	97.706	100
LH 725	Goeldi's marmoset	Primate	Z245	n/a	Potential novel	2	FNRW01000001.379172.380700_Bifidobacterium_longum	97.705	100
LH 389	Geoffroy's marmoset	Primate		n/a	Potential novel	3	AP012331.721637.723165_Bifidobacterium_scardovii_JCM_12489	95.176	100
LH 410	Red footed tortoise	Reptile		n/a	Potential novel	4	AWFT01000016.15.1546_Bifidobacterium_breve_MCC_1340	94.01	100
LH 418	Azara's agouti	Rodent		n/a	Potential novel	4	CP018044.576115.577667_Bifidobacterium_choerinum	94.444	100
LH 450	White-faced saki monkey	Primate		n/a	Potential novel	4	CP018044.576115.577667_Bifidobacterium_choerinum	94.444	100
LH 470	Azara's agouti	Rodent	Z241a	n/a	Potential novel	4	CP018044.576115.577667_Bifidobacterium_choerinum	94.972	100
LH 521	Azara's agouti	Rodent	Z241a	n/a	Potential novel	4	CP018044.576115.577667_Bifidobacterium_choerinum	95.531	99
LH 524	Azara's agouti	Rodent	Z241a	n/a	Potential novel	4	CP018044.576115.577667_Bifidobacterium_choerinum	95.531	100
LH 605	Azara's agouti	Rodent	Z211	n/a	Potential novel	4	CP018044.576115.577667_Bifidobacterium_choerinum	95.531	100
LH 716	Azara's agouti	Rodent	Z241b	n/a	Potential novel	4	CP018044.576115.577667_Bifidobacterium_choerinum	93.738	100
LH 717	Azara's agouti	Rodent	Z241b	n/a	Potential novel	4	CP018044.576115.577667_Bifidobacterium_choerinum	94.186	100
LH 718	Azara's agouti	Rodent	Z241b	n/a	Potential novel	4	CP018044.576115.577667_Bifidobacterium_choerinum	93.996	100
LH 413	Cotton top tamarin	Primate		n/a	Potential novel	5	16S sequence too short		
LH 419	Golden-headed lion tamarin	Primate		n/a	Potential novel	5	AP012331.721637.723165_Bifidobacterium_scardovii_JCM_12489	97.065	100
LH 424	Golden lion tamarin	Primate		n/a	Potential novel	5	AP012331.3080655.3082183_Bifidobacterium_scardovii_JCM_12489	96.924	100
LH 426	Goeldi's marmoset	Primate		n/a	Potential novel	5	AP012331.721637.723165_Bifidobacterium_scardovii_JCM_12489	96.464	100
LH 456	Emporer tamarin	Primate		n/a	Potential novel	5	AP012331.3080655.3082183_Bifidobacterium_scardovii_JCM_12489	96.924	100
LH 417	Cotton top tamarin	Primate		n/a	Potential novel	6	AP012331.721637.723165_Bifidobacterium_scardovii_JCM_12489	96.369	100
LH 421	Golden-headed lion tamarin	Primate		n/a	Potential novel	6	JGZP01000012.99390.100910_Bifidobacterium_stellenboschense	97.321	99
LH 423	Golden lion tamarin	Primate		n/a	Potential novel	7	AWFT01000016.15.1546_Bifidobacterium_breve_MCC_1340	97.186	100
LH 449	Scarlet ibis	Bird		n/a	Potential novel	8	CP007456.2269054.2270590_Bifidobacterium_kashiwanohense_PV20-2	95.901	100
LH 454	Geoffroy's marmoset	Primate		n/a	Potential novel	9	LKUR01000021.717.2255_Bifidobacterium_bifidum	95.416	100
LH 458	Emporer tamarin	Primate		n/a	Potential novel	10	CP001361.2165007.2166538_Bifidobacterium_bifidum_BGN4	95.366	100
LH 461	Pygmy marmoset	Primate		n/a	Potential novel	11	KC807989.1.1421_Bifidobacterium_aesculapii	99.502	99
LH 491	East African grey crowned crane	Bird		n/a	Potential novel	12	JGZJ01000010.292735.294257_Bifidobacterium_pullorum	96.595	99
LH 512	Abyssinian blue winged goose	Bird		n/a	Potential novel	13	CP021386.2234963.2236497_Bifidobacterium_breve	96.034	100
LH 558	Black lemur	Primate	Z174	n/a	Potential novel	14	JGZJ01000010.292735.294257_Bifidobacterium_pullorum	96.525	100
LH 560	Madagascan hissing cockroach	Insect	Z185	n/a	Potential novel	15	16S sequence too short		
LH 700	Madagascan hissing cockroach	Insect	Z271	n/a	Potential novel	15	AP012327.1791544.1793074_Bifidobacterium_kashiwanohense_JCM_15439	93.717	100
LH 702	Madagascan hissing cockroach	Insect	Z271	n/a	Potential novel	15	AP012327.1791544.1793074_Bifidobacterium_kashiwanohense_JCM_15439	93.717	100
LH 709	Madagascan hissing cockroach	Insect	Z271	n/a	Potential novel	15	AP012327.1791544.1793074_Bifidobacterium_kashiwanohense_JCM_15439	93.848	100
LH 687	Leopard tortoise	Reptile	Z267	n/a	Potential novel	16	KP718951.1.1509_Bifidobacterium_tissieri	94.685	99
LH 692	Leopard tortoise	Reptile	Z267	n/a	Potential novel	16	KP718951.1.1509_Bifidobacterium_tissieri	94.184	99
LH 838	White-faced duck	Bird	Z278	n/a	Potential novel	16	KP718951.1.1509_Bifidobacterium_tissieri	94.184	99
LH 935	Spoonbills	Bird	Z55/Z88	n/a	Potential novel	16	KP718951.1.1509_Bifidobacterium_tissieri	94.184	99
LH 730	Senegal bushbaby	Primate	Z260	n/a	Potential novel	17	AP010889.960839.962368_Bifidobacterium_infantis_ATCC_15697	96.141	100
LH 743	Golden-bellied mangabey	Primate	Z265	n/a	Potential novel	18	CP028341.2009205.2010740_Bifidobacterium_adolescentis	98.786	100
LH 898	White-faced owl	Bird	Z311	n/a	Potential novel	19	CP017696.1972737.1974271_Bifidobacterium_asteroides_DSM_20089	99.278	100

Table 5.1 Summary of the results of the ANI analysis and the screen of the 16S rRNA gene sequences predicted from the genomes of isolates identified to belong to putative novel Bifidobacterium species against SILVA 16S rRNA gene sequences database (v.138, 16 December 2019).
Strains displaying ANI values above 95% to each other were assigned to the same putative novel species.

Overall, the isolates predicted to belong to the same putative novel species showed highest 16S rRNA gene similarity values with a particular *Bifidobacterium* species, e.g. isolates predicted to belong to novel species '2' displayed highest 16S rRNA gene similarity values (above 97%) with strains of *B. longum*. Some exceptions to this observation included isolates predicted to belong to putative novel species '4' and '6'. For example, the majority of isolates assigned to species '4' showed highest 16S rRNA gene similarity to *B. choerinum*, with an exception of one tortoise-associated isolate which displayed highest 16S rRNA gene similarity to *B.* breve, with 94% similarity value. Overall, members of this group showed low predicted 16S rRNA gene sequence similarities (below 96%) to known *Bifidobacterium* species. These observations suggest that these isolates may be representative of a

new clade of *Bifidobacterium*. Another potential explanation of these findings may be the fact that the SILVA database does not yet contain sequences of most recently characterised animal-associated *Bifidobacterium* species that may be more closely related to putative novel isolates identified in this study.

To expand on the results of the 16S rRNA gene analysis, sequences for marker genes *rpoB, rpoC, groL, dnaJ, clpC* and *xpf* (*26*) were *in silico* extracted from annotated genomes of 87 *Bifidobacterium* type strains and the putative novel isolates, and concatenated for the purpose of multilocus sequence analysis. The quality trimmed multiple sequence alignment yielded 15,712 positions and was used for the phylogenetic tree reconstruction employing the maximum likelihood method (Figure 5.5).

The inclusion of genes from recently characterised *Bifidobacterium* species provided additional resolution for findings resulting from 16S rRNA gene analysis. Several of the isolates predicted to be sole representatives of putative novel species appeared to be more closely related to recently described *Bifidobacterium* species (e.g. LH_558 isolated from a lemur clustered with *Bifidobacterium lemurum* and *Bifidobacterium eulemuris* also derived from lemurs). Further assessment of phylogenetic relatedness between known members of the genus *Bifidobacterium* and the putative novel species indicated that a number of putative novel species comprising several isolates (e.g. species '2', '4', '15' and '16') may be representative of new clades. Notably, there appears to be a tendency for isolates obtained from the same host or environment to cluster by clade, but this clustering is not complete. Broadly, isolates recovered from primates seem to cluster with those obtained from humans, while insect-associated isolates appear to be more closely related to those obtained from birds and reptiles.

Figure 5.5 Phylogenetic tree based on concatenated housekeeping gene sequences for rpoB, rpoC, groL, dnaJ, clpC and xpf genes (15,712 nt).
The tree shows relationship between strains identified to belong to putative novel Bifidobacterium species and the type strains of currently recognised 87 Bifidobacterium species. The tree was reconstructed using the maximum likelihood method with 1000 bootstrap iterations and the sequences of Alloscardovia omnicolens DSM 21503^T, Alloscardovia omnicolens F0580 and Gardnerella vaginalis JCM 11026^T were used as an outgroup. Bootstrap values above 70% are marked on the tree branches. Coloured squares denote isolate distribution within putative novel Bifidobacterium species.

5.4.4 Genomic features

Previous studies suggested association between genomic signatures, e.g. genome size or specific genomic traits, and different bacterial habitats and lifestyles (*463, 474, 475*). To assess differences between *Bifidobacterium* isolates grouped within

host/environment categories, I examined their general genomic features and specific traits linked to their carbohydrate metabolism capabilities. For the isolates recovered from different hosts and environments, the range of genome sizes was from 1.63 Mb (*B. commune*) to 3.27 Mb (*B. myosotis*), corresponding to 1,357 and 2,628 protein-coding open reading frames, respectively. The genome sizes of human isolates ranged from 1.92 Mb to 3.16 Mb and corresponded to 1,606 and 2,593 ORFs, respectively. These values fall within the range for previously reported animal- and human-associated *Bifidobacterium* species (*33*).

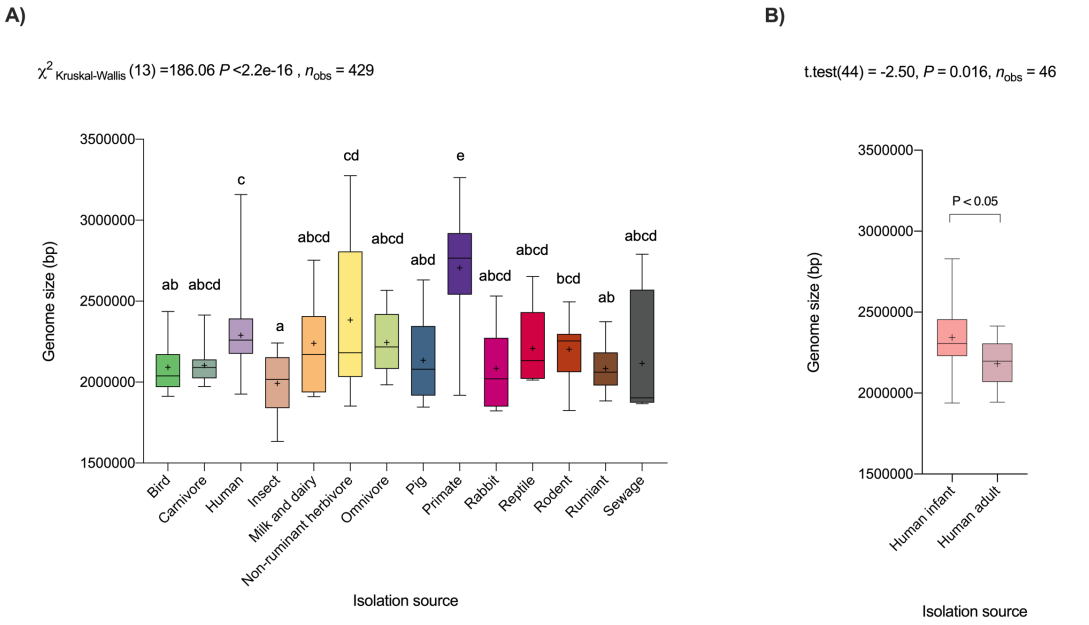


Figure 5.6 Genome sizes of Bifidobacterium isolates derived from (A) multiple hosts and environments and (B) humans.
The crosses represent the mean of each boxplot. The letters above each boxplot in the multiple host comparison diagram (A) represent the post hoc comparisons using Dunn's test, with groups sharing a letter not significantly different.

Comparison of genome sizes between groups indicated that isolates obtained from different hosts showed different trends with regard to genomes size (Kruskal-Wallis χ^2 = 186.06, P < 0.001, df = 13). Primate-associated *Bifidobacterium* isolates had the highest genome size (2.71 $\pm$ 0.30 Mb (mean $\pm$ sd)), while those obtained from insects displayed the lowest genome size (1.99 $\pm$ 0.18 Mb). These findings are in line with the results of the recent analysis of 129 publicly available *Bifidobacterium* strains (*31*). Similarly, the average genome size of isolates recovered from human

adults and infants displayed a significant difference (unpaired two-sample t-test, $P < 0.05$) (Figure 5.6).

To further assess the diversity among *Bifidobacterium*, I next sought to construct the pan-genome of 433 isolates and investigate phylogenetic relationships between them. To establish an informed gene identity threshold accounting for inter-species diversity, I first screened predicted proteomes of the isolates for the presence of marker genes based on the database of 105 conserved single copy genes defined by Dupont et al. (*476*). This analysis identified 72 marker genes shared by all *Bifidobacterium* isolates. The results of all-vs-all protein search of the custom database created from a subset of 10 identified marker gene sequences indicated relatively high diversity within the genus. For example, the minimum identity value for the marker gene *rnc* encoding ribonuclease III was 54.8% recorded between human-associated *B. animalis* subsp. *lactis* S7 and LH_439_A isolated from a probiotic formulation (Table 5.2). Based on these findings, I set the protein search threshold for pan-genome construction at 50%.

Minimum identity values between marker genes (blastp, evalue 1e-50)

Gene	Description	Identity% (coverage%)	Isolate
rpoB	DNA-directed RNA polymerase, beta subunit	82.1 (100)	B. catulorum MRM8_19 -> B. thermacidophilum subsp. thermacidophilum LMG 21395
rpoC	DNA-directed RNA polymerase, beta' or beta" subunit	76.8 (99)	B. callitrichos UMA 51804 -> LH_687
clpC	ATP-dependent Clp protease	78.6 (98)	B. bombi DSM 19703 -> LH_687
rplJ	ribosomal protein L10	68.8 (100)	B. dentium 1893B -> LH_709
smpB	SmpB protein	71.6 (99)	B. anseris Goo31D -> LH_935
serS	seryl-tRNA synthetase	72.8 (99)	B. bohemicum DSM 22767 -> LH_912
rpsB	ribosomal protein S2	69.4 (100)	B. coryneforme DSM 20216 -> LH_543
rnc	ribonuclease III	54.8 (93)	B. animalis subsp. lactis S7 -> LH_439_A
pgk	phosphoglycerate kinase	76.1 (100)	B. asteroides ESL0200 -> B. catulorum MRM8_19
rplE	ribosomal protein L5	78.4 (98)	B. criceti Ham19E -> LH_709

Table 5.2 Minimum identity values for the all-vs-all BLASTP comparison (e-value 1e-50) between selected marker genes shared by the 433 Bifidobacterium isolates.

The results of this analysis identified a total of 46,766 clusters of orthologous genes, representing the pan-genome of the 433 *Bifidobacterium* isolates (Figure 5.7). Functional classification of the overall gene content resulted in the assignment of functional categories to 76.9% of ORFs, with the remaining 23.1% categorised as proteins of unknown function.

A) B)

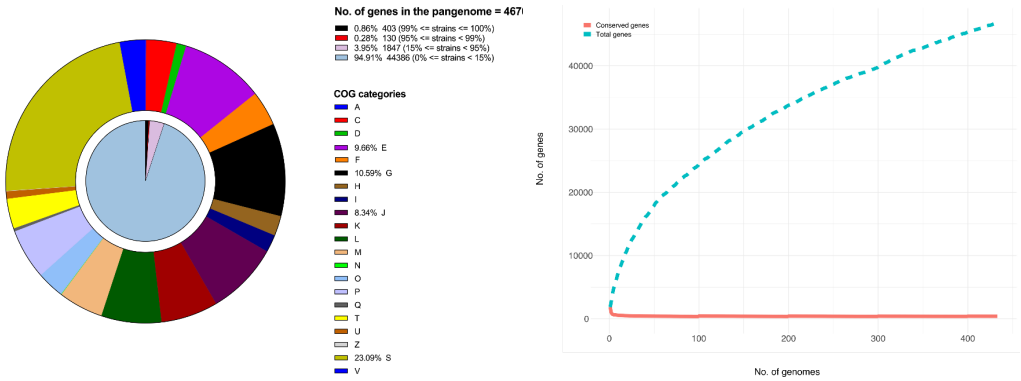


Figure 5.7 Pan-genome of the genus Bifidobacterium.
(A) The internal pie chart summarises the pan-genome of 433 isolates and shows the core genes (n=403, black), the soft core genes (n=130, red), the shell genes (n=1,847, light purple) and the cloud genes (n=44,386, light blue) determined based on the default Roary thresholds. The external pie chart represents the COG classification of the whole Bifidobacterium pan-genome, with specific categories highlighted in different colours. The values for the most abundant categories are presented in the diagram. Summary of COG categories identified in the Bifidobacterium pan-genome: [A] RNA processing and modification; [C] Energy production and conversion; [D] Cell cycle control, cell division, chromosome partitioning; [E] Amino acid transport and metabolism; [F] Nucleotide transport and metabolism; [G] Carbohydrate transport and metabolism; [H] Coenzyme transport and metabolism; [I] Lipid transport and metabolism; [J] Translation, ribosomal structure and biogenesis; [K] Transcription; [L] Replication, recombination and repair; [M] Cell wall/membrane/envelope biogenesis; [N] Cell motility; [O] Post-translational modification, protein turnover, and chaperones; [P] Inorganic ion transport and metabolism; [Q] Secondary metabolites biosynthesis, transport, and catabolism; [S] Function unknown; [T] Signal transduction mechanisms; [U] Intracellular trafficking, secretion, and vesicular transport; [V] Defence mechanisms; [Z] Cytoskeleton.
(B) Pan-genome (dashed teal line) and core genome (solid red line) of the genus Bifidobacterium. The pan-genome and core genome are represented as sizes of their gene pools versus the analysed 433 bifidobacterial genomes. The x axis represents the number of genomes, whereas the y axis represents the number of genes.

Carbohydrate transport and metabolism was identified as second most abundant category and constituted 10.6% of *Bifidobacterium* pan-genome, reflecting the saccharolytic lifestyle of members of this genus. This value is within the range previously reported for bifidobacteria (*162, 293*) (Figure 5.7 & Table S5.4). Other abundant functional categories included genes involved in amino acid metabolism (9.66%) and genes related to translation, ribosomal structure and biogenesis (8.34%).

Plotting the pan-genome size versus the number of genomes included in the analysis showed that the power trendline has not reached a plateau, indicating that

the addition of new genome sequences to the pan-genome would still increase the total gene pool. This suggest that there is scope for further discovery of diversity within the *Bifidobacterium* genus.

Notably, 403 genes were shared among all the isolates, constituting the core genome. These findings resonate with observations from previous studies with fewer genomes, which reported the number of genes in the core genome of genus *Bifidobacterium* to range between 400 and 500 genes (*27, 29, 31*). Plotting the number of identified core genes as a function of the total number of genomes included in the analysis showed that the trendline had reached a plateau, suggesting that the addition of further genome sequences to the pan-genome would not reduce the core genome of the genus significantly.

Inclusion of genome sequences of other members of the family *Bifidobacteriaceae*, namely *Alloscardovia omnicolens* DSM 21503^T, *Alloscardovia omnicolens* F0580 and *Gardnerella vaginalis* JCM 11026^T, generated a set of 392 core genes shared by all included isolates. Concatenated sequences for these genes were used to construct the family phylogeny (Figure 5.8). The structure of the generated phylogenomic tree was similar to previously reported phylogenies generated based on the 16S rRNA gene and various sets of core genes (*25, 286*), and clearly delineated the presence of seven previously described bifidobacterial phylogenetic groups, i.e. *B. adolescentis*, *B. asteroides*, *B. bifidum*, *B. boum*, *B. longum*, *B. pseudolongum*, and *B. pullorum*. The members of the same taxonomic species clustered together. In contrast to previous observations positioning the *B. asteroides* phylogroup in the deepest branches of the lineage, the deepest branches of the phylogeny corresponded to isolates belonging to *Bifidobacterium tsurumiense* species and the bird-associated strain LH_491, representative of a potential novel *Bifidobacterium* species.

Comparison between the phylogenetic relatedness and the host groups indicated that there was a statistically significant association between the isolation source and the phylogenetic distribution of *Bifidobacterium* (ANOSIM: R = 0.4016, $P < 0.001$). In particular, strains derived from primates, insects and rodents seemed to cluster closely within their respective phylogroups. This clustering was not

perfect however, with clades of strains from mixed isolation sources present within distinct phylogenetic groups in the tree (e.g. in the *B. pseudolongum* phylogroup). In these cases, the tendency for strains recovered from particular host groups to cluster together was more apparent at the higher taxonomic level of the host, for example for orders Artiodactyla (strains isolated from ruminants, non-ruminant herbivores and pigs) or Primates (strains derived from humans and non-human primates).

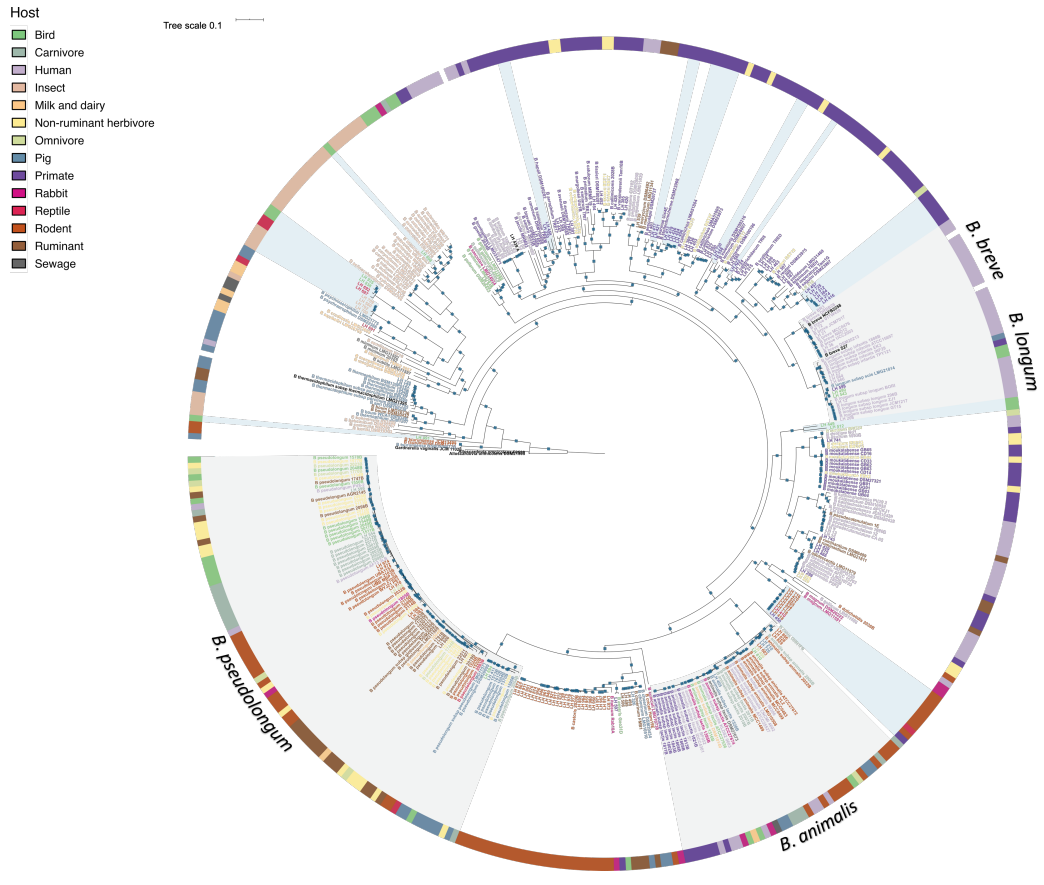


Figure 5.8 Phylogenomic overview of the genus Bifidobacterium.
Maximum likelihood phylogeny was based on concatenated sequences for 392 core genes, employing the 'GTR' model with 1000 bootstrap iterations. Bootstrap values above 70% are displayed on tree branches. Grey shading depicts the prominent recognised Bifidobacterium species, while blue shading marks isolates identified in this study as belonging to putative novel Bifidobacterium species. The sequences of Alloscardovia omnicolens DSM 21503^T, Alloscardovia omnicolens F0580 and Gardnerella vaginalis JCM 11026^T were used as an outgroup.

5.4.5 Glycobiome

Given that carbohydrate metabolism and transport was the second most abundant functional group identified in the *Bifidobacterium* pan-genome, I next sought to investigate the genetic repertoire predicted to be involved in carbohydrate processing. *In silico* analysis performed using dbCAN2 identified carbohydrate-active enzymes (CAZymes) in genomes of analysed isolates (Table S5.5). Previous findings on CAZyme abundances in strains isolated from different hosts and environments indicated that, on average, *Bifidobacterium* isolated from non-human primates encoded the highest number of CAZymes (84 genes ± 20 sd), whereas those recovered from wastewater carried the fewest genes associated with carbohydrate metabolism (42 genes ± 10 sd) (*31*). The predictions generated for 433 *Bifidobacterium* isolates were in line with these observations, with average numbers of predicted CAZymes overall higher across host groups (103.81 genes ± 23.09 sd for primate category and 57.00 genes ± 11.63 sd for sewage, respectively).

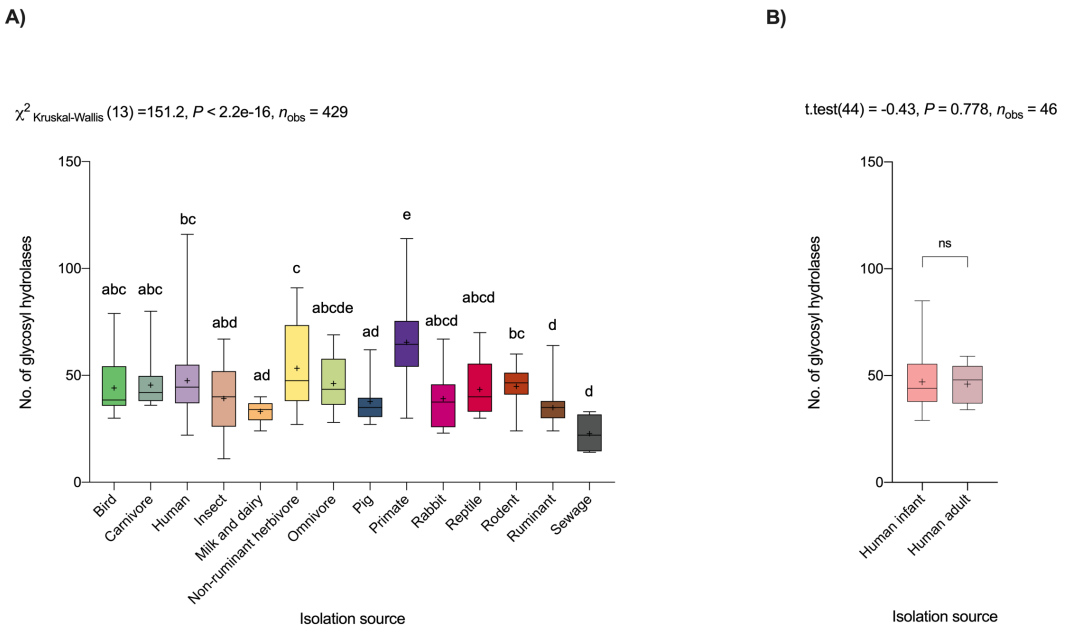


Figure 5.9 Abundance of glycosyl hydrolase families in isolates derived from (A) multiple hosts and environments and (B) humans.
The crosses represent the mean of each boxplot. The letters above each boxplot in the multiple host comparison diagram (A) represent the post hoc comparisons using Dunn's test, with groups sharing a letter not significantly different.

Furthermore, the results indicated that over 50% of CAZymes in each host category belong to glycosyl hydrolase class of enzymes (range between 51% and 63%), with an exception of sewage; in this group GHs constituted 40% of predicted carbohydrate-active enzymes (Table S5.6). Comparison of glycosyl hydrolase numbers across different host groups revealed significant differences in GH abundance between categories (Kruskal-Wallis $\chi^2 = 151.2$, $P < 0.001$, df = 13) (Figure 5.9). Isolates belonging to the primate group carried the highest average number of GH genes (65.68 genes $\pm$ 19.17 sd), while those associated with sewage harboured the fewest (22.75 genes $\pm$ 9.21 sd). In contrast, the abundance of GHs between human-derived infant and adult strains did not differ significantly (unpaired two-sample t-test, $P > 0.05$). On average, adult isolates harboured slightly fewer glycosyl hydrolase genes (46.00 genes $\pm$ 9.58 sd) compared to infant-associated *Bifidobacterium* (47.09 genes $\pm$ 11.08 sd).

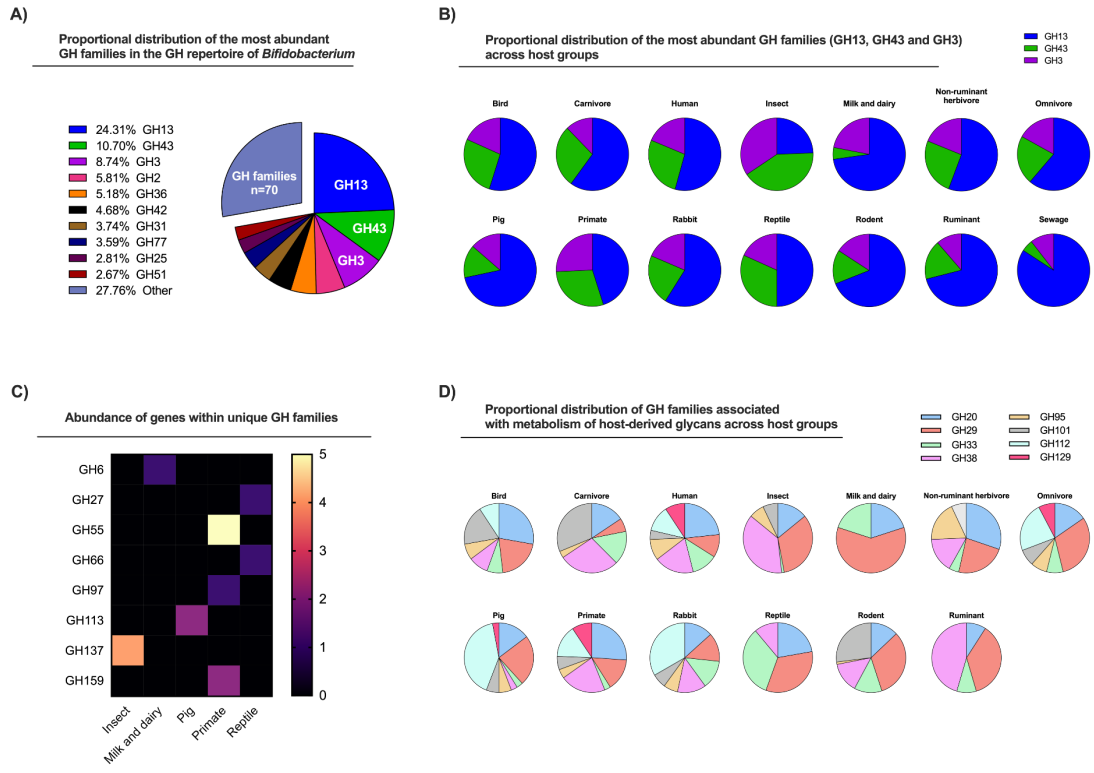


Figure 5.10 Distribution of selected GH families across Bifidobacterium isolates. (A) Proportional distribution of the most abundant GH families in the GH repertoire of Bifidobacterium. (B) Proportional distribution of the three most abundant GH families across host groups. (C) Abundance of genes within unique GH families identified across host groups. (D) Proportional distribution of GH families associated with metabolism of host-derived glycans across host groups.

A total of 80 GH families were identified across the 433 members of the genus *Bifidobacterium*. GH13 constituted the predominant GH family of the genus (24.3%) (Figure 5.10 A & B). As stated in Chapters 3 and 4, enzymes belonging to this family hydrolyse a wide range of complex carbohydrates, including starch, glycogen and related substrates, as well as stachyose, raffinose, palatinose, and melibiose. The complete breakdown of the latter group of carbohydrates has been shown to involve enzymes belonging to the GH36 family, which was also ubiquitous across the analysed host categories (*421*). On average, *Bifidobacterium* strains harboured 11.86±3.51 GH13 genes per genome, with an exception of insect-associated isolates, whose genomes contained 3.21±1.61 predicted GH13 genes (Table S5.6). Other major ubiquitous GH families included GH43 and GH3, both associated with the degradation of plant-derived carbohydrates, with strains harbouring an average of 5.23±4.44 and 4.27±3.26 GH genes per genome, respectively. These observations are in line with previous findings resulting from the analysis of 47 *Bifidobacterium* type strains (*38*).

Eight GH families were unique to subsets of isolates in particular host groups (Figure 5.10 C). Further examination of these results indicated that the unique GH families with the highest abundance of GH genes, namely GH55 (exo- and endo-β-1,3-glucanases, n=5 genes) and GH137 (β-L-arabinofuranosidases, n=4 genes), were predicted in primate- and insect-associated isolates identified as members of two distinct potentially novel *Bifidobacterium* species (species '5' and species '15', respectively) (Table S5.6).

As stated previously, various species and strains of *Bifidobacterium* have been shown to possess enzymes involved in the degradation of host-associated glycans, e.g. milk oligosaccharides and mucin. Examples of such enzymes include members of families GH33 (exo-sialidases), GH29 and GH95 (α-fucosidases), GH20 (enzymes with hexosaminidase and lacto-*N*-biosidase activities), GH112 (lacto-*N*-biosidases), GH38 and GH125 (α-mannosidases), as well as GH101 and GH129 (α-*N*-acetylgalactosaminidases) (*38*). Evaluation of the presence of these GH families across the host groups revealed that sewage-associated isolates did not harbour

any of the genes associated with the metabolism of host-derived carbohydrates (Figure 5.10 D). Furthermore, dairy-, reptile- and ruminant-associated isolates possessed a more limited repertoire of host glycan-associated GH families compared to the remaining host groups, with members of GH95, GH101, GH112 and GH129 absent from their genomes. Overall, these results illustrate the potential of members of the genus for broad carbohydrate metabolism capabilities.

5.4.6 Discussion

This work provides new insights into the diversity of a substantial animal-associated *Bifidobacterium* sample set and associated genomic dataset. Over 100 unduplicated *Bifidobacterium* isolates were recovered from a range of animal hosts, including previously unreported species of rodents (Azara's agouti), reptiles (tortoises) and insects (Madagascan hissing cockroach). Genomic-based approaches, namely the ANI analysis and the multilocus sequence analysis, primarily identified 44 isolates as members of 19 putative novel *Bifidobacterium* species.

Previous studies have suggested a link between bacterial genome size and the habitats occupied by organisms, exploring the concept that particular habitats harbour a specific range of genome sizes, which are related to factors intrinsic to these given habitats (*463, 474, 475*). Metagenomic studies of bacterial communities correlated average genome length with habitat dynamicity, according to the hypothesis that organisms with larger genomes can thrive in complex habitats by encoding a wider repertoire of proteins with functions related to metabolism and stress tolerance (*477-479*). Environmental pressures acting on organisms over time, coupled with vertical inheritance of genes through phylogenetic descent, have been named as factors shaping the content of microbial genomes (*474, 480*). One outcome of such processes is the genome reduction in bacteria adapted to symbiotic lifestyles (*463, 481, 482*), described for example for insect-associated obligate symbionts (*483, 484*).

Comparison of bifidobacterial genomic signatures across the host groups revealed significant differences between several host categories in terms of genome size and abundance of genes involved in carbohydrate metabolism. In particular, genomes of strains associated with non-human primates were significantly larger than any other group. These strains also harboured the highest number of predicted CAZymes, and consequently, the highest number of genes encoding glycosyl hydrolases. These results may be explained in part, by the varied plant diets of primates. Based on the detailed dietary information obtained from collaborators at the Banham Zoo, the diet of *Bifidobacterium*-positive primates consisted of, alongside protein sources, a large variety of fruit and vegetables, Arabic gum and commercial dietary formulations for primates containing various complex carbohydrates, such as pectins, hemicelluloses, celulloses, lignins and starches. Additionally, complexity and diversity of primate milk oligosaccharides may contribute to the increased repertoire of genes involved in the degradation of complex carbohydrates in this group (*485*). The search for the presence of GH families previously linked to metabolism of host-derived glycans in humans (e.g. milk oligosaccharides and mucin), revealed that primate-associated isolates possessed a repertoire of these GH families similar to that of human-associated *Bifidobacterium* (Figure 5.10) (*295*).

In contrast, insect-associated *Bifidobacterium* isolates displayed the smallest genome size and relatively low abundance of glycosyl hydrolases. In particular, their genomes were predicted to encode a very limited repertoire of genes belonging to GH13, compared to *Bifidobacterium* isolates associated with other host groups. These observations reflect the findings from previous studies, which suggested that the low abundance of GH13 genes in insect-associated *Bifidobacterium* type strains recovered from honey bees and bumblebees may be linked to their particular diet essentially lacking carbohydrates with α-glucoside linkages (nectar and pollen) (*38, 486*). When compared to bee-associated isolates, those recovered from cockroaches had a higher average number of GH13 genes in their genomes, 6.00 genes $\pm$ 0.00 sd vs. 2.76 genes $\pm$ 1.23 sd. These findings may result from differences in the diet between the insect hosts. According to the metadata obtained from the

Zoo, the cockroach diet consisted of fresh fruit, vegetables and leaf litter, likely to contain carbohydrates prone to the action of GH13 enzymes.

Comparison within the human category revealed that genomes of isolates associated with infants appeared to be significantly larger, on average, than those of isolates recovered from adults. This observation conflicts with the work of Rodriguez and Martiny (*31*), who did not find differences between these groups. These opposing results may be explained by the fact that the data subset analysed here contained an overall higher number of human-associated isolates (n=46 [n=34 infant and n=12 adult] compared to n=20 [n=10 infant and n=10 adult]) and encompassed higher strain diversity, including a higher number of *B. longum* subsp. *infantis* isolates, which are characterised by larger genome sizes compared to other species associated with infant hosts. In line with previous findings (*31*), comparison of the abundance of GH families between infant- and adult-associated isolates did not reveal significant differences.

Exploring similarities between host groups and across the phylogenetic tree revealed that *Bifidobacterium* isolates recovered from the same host or environment were non-randomly associated with their phylogenetic distribution. These results are consistent with previous findings form studies with fewer genomes (*31*). It has been hypothesized that certain species of *Bifidobacterium* prefer particular hosts, while others exhibit more a more “cosmopolitan” host distribution (*487*). Several studies found that isolates associated with the same host, or a specific group of hosts, tended to form phylogenetic subclusters (*31, 293, 487*). While this pattern was present for some of the species represented by fewer strains, e.g. *B. ruminantium* or *B. boum*, many other isolates from the same host could be found within several clades and subclusters (e.g. *B. animalis*). These results suggested that bifidobacteria may not exhibit strict host specialisation, however the analysis of a larger dataset encompassing larger sample sizes and a wider host range would allow for more robust observations.

Several studies have analysed the *Bifidobacterium* pan-genome, with the number of included isolates ranging from 14 to 215 (*27-31*), with all studies suggesting an open pan-genome. Doubling the number of genomes (n = 433) in this analysis

indicated the pan-genome is still open, which suggest a continuous *Bifidobacterium* expansion, potentially driven by a range of evolutionary mechanisms, such as HGT, speciation and evolutionary selection pressure acting on members of the genus (*30*). Gene gain and loss events, corresponding to changes in bacterial physiology and lifestyle, shape the gene repertoire of species and reflect their ability to adapt to a wide range of ecological niches and to respond to the environmental pressures (*488*). Given that the 433 *Bifidobacterium* isolates inhabit a wide range of hosts, and based on the results of this preliminary analysis, I expect that significant inter- and intra-species variations exist within the genus, leading to niche-specific genomic adaptations. Future work would test this hypothesis.

5.5 Future work

Through a combination of bacterial isolation, sequencing, and bioinformatic analysis, this study sought to provide additional insights into the diversity of animal-associated members of the genus *Bifidobacterium*. The isolation efforts resulted in the recovery of a substantial number of isolates from different hosts and the identification of strains representative of almost 20 putative novel *Bifidobacterium* species. Candidates for biochemical characterisation as type strains would be selected from the latter pool of isolates, and complementary bioinformatic analyses would be performed according to the proposed standards for the use of genome sequences for the taxonomic classification of bacteria (*308*). As part of the zoo microbiome project, all animal faecal samples subjected to the isolation experiments were also processed for 16S rRNA gene sequencing, and this data would be analysed to gain insight into the environment the *Bifidobacterium* isolates were recovered from (i.e. wider bacterial community), and assess the relative abundance of members of this genus in different animal hosts.

Although this analysis represents the largest to date, preliminary observations indicate that the gene content of the genus has not yet been resolved (open pan-genome), and there are significant differences in genomic signatures (e.g. genome size, number of GHs) between isolates associated with specific host

categories. More detailed examination of the functional aspects of the pan-genome would allow to further assessment of inter- and intra-species variation within and across host categories. Finer-scale evolutionary mechanisms, including for example glycosyl hydrolase gene gain and loss events or horizontal gene transfer events, would be explored through further strain-level bioinformatic analyses. Based on the results of these analyses, a subset of isolates displaying interesting (e.g. health-promoting or previously unreported) features could be selected, and experimental validation of the computational results carried out. For example, the assessment of carbohydrate metabolism profiles of selected isolates and their co-operation with other members of the gut bacterial community could result in findings leading to potential application in the veterinary or animal husbandry industry.

Broader assessments of the diversity of *Bifidobacterium* and the complete resolution of their pan-genome would not be possible without the inclusion of additional isolates. Currently, the sampling among host animals is quite uneven, and certain host categories are significantly underrepresented (e.g. insects and reptiles). In addition, data on *Bifidobacterium* distribution and prevalence in wild animals are lacking. Further isolation efforts from larger sample sizes among a broader range of hosts, e.g. birds, reptiles and insects, and the subsequent generation of genomic datasets would strengthen the results of diversity analyses. Additionally, targeting as yet uncultured species and strains of *Bifidobacterium* for genome sequencing would improve the representation of the true diversity of the genus in their habitats.

Chapter 6

Final considerations

Bifidobacterium are key members of the gut microbiota and are widely distributed in the animal kingdom (*6*). While the true diversity of animal-associated *Bifidobacterium* has yet to be evaluated, human-associated members of this genus predominate the gut microbiota during early life, with many strains capable of metabolising breast milk-derived HMOs (*133*). *Bifidobacterium* from later life points have been shown to carry genes for the degradation of plant-derived complex carbohydrates (*33*). Overall, in this thesis I have shown an association between host diet and *Bifidobacterium* species and strains in both humans and animals. Bioinformatic and experimental approaches revealed genomic flexibility and growth plasticity within *Bifidobacterium*, which may be the feature that allows colonisation of a wide range of hosts and is linked to the ability of particular species and strains to metabolise distinct dietary components.

The high abundance of *Bifidobacterium* in early infancy, and strains of *B. longum* in particular, has been linked to nutrient availability, however longitudinal assessments of these bacteria during the crucial transition from milk-based diet to solid foods have been lacking. This aspect has been explored as part of my PhD through the genomic analysis of 75 *B. longum* strains isolated from nine individual infants at different dietary stages, i.e. pre-weaning, weaning and post-weaning, with a particular focus on their potential carbohydrate utilisation capabilities. The analysed strains showed intra-individual and diet-related differences in their genomic content, which could be associated with their ability to degrade specific dietary components.

Genomic analysis revealed genome flexibility within *B. longum*, with intra-subspecies differences in GH family content between strains. *B. infantis* was found to predominantly contain GH families implicated in the fermentation of host-derived glycans like HMOs, which property could confer a particular advantage for this subspecies and facilitate their establishment in the early life microbiota. In

contrast, *B. longum* predominantly contained GH families involved in the degradation of plant-derived substrates, which could influence their ability of to persist within individual hosts. Indeed, I observed that particular strains of *B. longum* could persist in individuals through infancy, for at least 18 months, despite significant changes in diet. At the same time new strains displaying different genomic content and potential carbohydrate metabolism capabilities appeared be acquired, possibly in response to the changing nutritional environment. These results were further supported by subspecies-specific differences in GH content between pre- and post-weaning strains, and across the *B. longum* subspecies between strains isolated from breast-fed and formula-fed strains at different dietary stages.

The complementary phenotypic growth data suggested that *Bifidobacterium* possess an overall very broad repertoire of carbohydrate utilisation genes that may be differentially switched on and off in response to the presence of specific dietary components (*392, 393*) This was reflected in differential growth of genotypically similar *B. longum* strains on various carbohydrate sources. The experimental data from formula-fed isolates were particularly interesting, with several closely related weaning and post-weaning strains from one baby able to metabolise selected HMOs. All babies recruited for this study only received standard formula not supplemented with any prebiotics or synthetic HMOs.

Collectively, results from genomic and phenotypic analyses indicated a strong association between host diet and *Bifidobacterium* species and strains, which seemed to align with changes to the nutritional environment during weaning. However, the assessment of *B. infantis*, in particular in formula-fed babies, requires further efforts. Limited availability of isolates made it impossible to examine properties of *B. infantis* within this dietary group and make comparisons with breast-fed strains. Although *B. infantis* is primarily associated with the degradation of HMOs in breast-fed infants, I recorded growth of one of the *B. infantis* strains from formula-fed baby on xylose, albeit with inconsistent outcomes between experiments. Further experimental assays using a larger number of isolates would

be required to fully assess the ability of this subspecies to degrade a wider range of non-HMO carbohydrates in early life.

Human-associated bifidobacteria have received much attention in the recent years due to their probiotic properties and associated beneficial health outcomes for the host. However, members of *Bifidobacterium* have a wide host range and have been extensively detected in the gut of non-human mammals, birds and social insects (*6*). Despite an increase in the availability of genomic data, studies focusing on members of *Bifidobacterium* in animals have been lacking, with the pan-genome and phylogenetic relationship of the genus unresolved. The sampling among host animals is quite uneven, and certain host categories are significantly underrepresented (e.g. insects and reptiles). An important aspect of this PhD project was the isolation of *Bifidobacterium* from a range of animal hosts, with pioneering examination of members of this genus in wild mice populations. These efforts, coupled with sequencing and bioinformatic analysis, resulted in the recovery of a substantial number of isolates (over 100) from various animals, including previously unreported species of rodents, reptiles and insects, and in the identification of strains representative of almost 20 putative novel *Bifidobacterium* species.

Global examination of over 400 *Bifidobacterium* genomes, including publicly available sequences, revealed significant differences in genomic signatures (e.g. genome size and abundance of glycosyl hydrolases) between isolates associated with different host groups, supporting the notion that particular habitats harbour a specific range of genome sizes, which are related to factors intrinsic to these given habitats (*463, 474, 475*). Differences in the presence and the abundance of particular GH families in various host groups could be explained in part by host diet. Presence of high numbers of GH families associated with degradation of plant-derived carbohydrates in primate-associated *Bifidobacterium* isolates could be linked to varied plant components consumed by primates. Similarly, the identification of previously unreported GH46 family containing chitosanases in genomes of *B. castoris* associated with wild rodents could

potentially be explained by the presence of fungal component in the host diet (*443*).

The exploration of similarities between host groups and across the phylogenetic tree revealed non-randomly association between *Bifidobacterium* isolates recovered from the same host or environment and their phylogenetic distribution. It has previously been proposed that certain species of *Bifidobacterium* prefer particular hosts, while others exhibit a more global host distribution (*487*), and that isolates associated with the same host, or a specific group of hosts, tended to form phylogenetic subclusters (*31, 293, 487*). This pattern was observed in this work for some of the *Bifidobacterium* species represented by fewer strains, but other isolates from the same host could be found within several clades and subclusters in the tree, suggesting that bifidobacteria may not exhibit strict host specialisation. Interestingly, to date all *B. castoris* isolates have been recovered from rodents. Additional isolation or high-resolution metagenomic efforts would be required to establish whether this species is restricted to this particular host group. Overall, the analysis of a larger dataset encompassing larger sample sizes and a wider host range would allow for more robust observations on *Bifidobacterium* specificity.

The pan-genomic analysis in this work represents the largest to date, with preliminary observations indicating that the gene content of the genus has not yet been resolved and suggesting an open pan-genome. These results are in line with the findings from previous studies with fewer genomes (*27-31*), and are an indication of a continuous *Bifidobacterium* expansion, potentially driven by a range of evolutionary mechanisms, such as horizontal gene transfer events, speciation and evolutionary selection pressure acting on members of the genus (*30*). Significant inter- and intra-species variations leading to niche-specific genomic adaptations are expected to exist within the genus. More detailed examination of the functional aspects of the *Bifidobacterium* pan-genome, coupled with the bioinformatic exploration of finer-scale evolutionary mechanisms, such as glycosyl hydrolase gene gain and loss events or HGT events, would allow further assessment of this expected diversity within and across host groups.

Subsequent experimental validation of isolates presenting health-promoting or previously unreported properties could reveal potential for industrial applications, for example in animal husbandry or captive animal welfare. The identification of predicted GH families GH46 and GH49 in *B. castoris*, and families GH55 and GH137 in putative novel *Bifidobacterium* species isolated from primates and insects, respectively, offers prospect for further investigation. Experimental confirmation of functionality coupled with transcriptomic and proteomic analysis could lay foundation for the characterisation of catalytic activity of recombinant enzymes.

Potentially, Type II CRISPR-Cas systems can be exploited to create new tools for genome editing and engineering. Additional bioinformatic investigations could reveal the presence of CRISPR-Cas systems in newly isolated animal-associated *Bifidobacterium*. The identification of the complete Type II CRISPR-Cas system in *B. castoris* is promising and could be followed by bioinformatic analysis of protospacer adjacent motifs (PAMs) and further experimental approaches to investigate whether the identified CRISPR-Cas systems are actively transcribed and protect against invasive DNA, e.g. prophages.

The results of the investigation of prophages in *B. castoris* genomes could be further strengthened by integration of data from other available software to improve the identification of prophage-associated genes. A similar approach could be used to screen for prophage elements in the genomes of newly isolated *Bifidobacterium*. These data could then be explored to assess whether experimental prophage induction is possible. Successful phage isolation would significantly expand the currently very limited data on *Bifidobacterium*-associated phages.

Overall, this PhD research has contributed significantly to the current understanding of genomic and phenotypic properties of *Bifidobacterium.* These data should lay the foundation for future in-depth research aiming at the assessment of the extensive diversity of animal-associated members of this genus and their functional potential for both therapeutic and industrial applications.

References

1. W. B. Whitman *et al.*, *Bergey's manual of systematic bacteriology. Volume 5: The Actinobacteria, part A and B*. (Springer, New York, ed. 2nd, 2012), pp. 2083.
2. H. Tissier, *Recherches sur la flore intestinale des nourrissons (état normal et pathologique)*. (Paris, 1900), pp. 253.
3. D. H. Bergey, R. E. Buchanan, N. E. Gibbons, *Bergey's manual of determinative bacteriology*. (Williams & Wilkins Co., Baltimore, ed. 8th, 1974), pp. 1268.
4. W. de Vries, A. H. Stouthamer, Pathway of glucose fermentation in relation to the taxonomy of bifidobacteria. *J. Bacteriol.* **93**, 574-576 (1967).
5. E. W. Sayers *et al.*, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **48**, D9-D16 (2020).
6. F. Turroni, D. van Sinderen, M. Ventura, Genomics and ecological overview of the genus Bifidobacterium. *Int. J. Food Microbiol.* **149**, 37-44 (2011).
7. J. L. Rasic, J. A. Kurmann, Bifidobacteria and their role. Microbiological, nutritional-physiological, medical and technological aspects and bibliography. *Experientia Suppl* **39**, 1-295 (1983).
8. R. Hartemink, F. M. Rombouts, Comparison of media for the detection of bifidobacteria, lactobacilli and total anaerobes from faecal samples. *J. Microbiol. Methods* **36**, 181-192 (1999).
9. D. Roy, Media for the isolation and enumeration of bifidobacteria in dairy products. *Int. J. Food Microbiol.* **69**, 167-182 (2001).
10. H. Beerens, Detection of bifidobacteria by using propionic acid as a selective agent. *Appl. Environ. Microbiol.* **57**, 2418-2419 (1991).
11. V. Rada, J. Koc, The use of mupirocin for selective enumeration of bifidobacteria in fermented milk products. *Milchwissenschaft* **55**, 65-67 (2000).
12. P. Mattarelli *et al.*, Recommended minimal standards for description of new taxa of the genera Bifidobacterium, Lactobacillus and related genera. *Int. J. Syst. Evol. Microbiol.* **64**, 1434-1451 (2014).
13. K. T. Konstantinidis, J. M. Tiedje, Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr. Opin. Microbiol.* **10**, 504-509 (2007).
14. J. Chun, F. A. Rainey, Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. *Int. J. Syst. Evol. Microbiol.* **64**, 316-324 (2014).
15. O. Mizrahi-Man, E. R. Davenport, Y. Gilad, Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. *Plos One* **8**, e53608 (2013).
16. E. Stackebrandt, Goebel, B. M., Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *Int. J. Syst. Evol. Microbiol.* **44**, 846-849 (1994).
17. P. P. Bosshard *et al.*, 16S rRNA gene sequencing versus the API 20 NE system and the VITEK 2 ID-GNB card for identification of nonfermenting Gram-negative bacteria in the clinical laboratory. *J. Clin. Microbiol.* **44**, 1359-1366 (2006).
18. S. Mignard, J. P. Flandrois, 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. *J. Microbiol. Methods* **67**, 574-581 (2006).
19. M. Kim, H. S. Oh, S. C. Park, J. Chun, Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int. J. Syst. Evol. Microbiol.* **64**, 346-351 (2014).
20. I. Olsen, J. L. Johnson, L. V. H. Moore, W. E. C. Moore, Rejection of Clostridium putrificum and conservation of Clostridium botulinum and Clostridium sporogenes-

Opinion 69. Judicial Commission of the International Committee on Systematic Bacteriology. *Int. J. Syst. Bacteriol.* **49 Pt 1**, 339 (1995).
21. P. E. Fournier, D. Raoult, Current knowledge on phylogeny and taxonomy of Rickettsia spp. *Ann. N. Y. Acad. Sci.* **1166**, 1-11 (2009).
22. A. Roth, S. Andrees, R. M. Kroppenstedt, D. Harmsen, H. Mauch, Phylogeny of the genus Nocardia based on reassessed 16S rRNA gene sequences reveals underspeciation and division of strains classified as Nocardia asteroides into three established species and two unnamed taxons. *J. Clin. Microbiol.* **41**, 851-856 (2003).
23. M. Rossi-Tamisier, S. Benamar, D. Raoult, P. E. Fournier, Cautionary tale of using 16S rRNA gene sequence similarity values in identification of human-associated bacterial species. *Int. J. Syst. Evol. Microbiol.* **65**, 1929-1934 (2015).
24. F. Turroni *et al.*, Exploring the Diversity of the Bifidobacterial Population in the Human Intestinal Tract. *Appl. Environ. Microbiol.* **75**, 1534-1545 (2009).
25. G. A. Lugli *et al.*, Investigation of the Evolutionary Development of the Genus Bifidobacterium by Comparative Genomics. *Appl. Environ. Microbiol.* **80**, 6383-6394 (2014).
26. M. Ventura *et al.*, Analysis of bifidobacterial evolution using a multilocus approach. *Int. J. Syst. Evol. Microbiol.* **56**, 2783-2792 (2006).
27. C. Milani *et al.*, Genomic Encyclopedia of Type Strains of the Genus Bifidobacterium. *Appl. Environ. Microbiol.* **80**, 6290-6302 (2014).
28. F. Bottacini *et al.*, Comparative genomics of the genus Bifidobacterium. *Microbiol-Sgm* **156**, 3243-3254 (2010).
29. Z. Sun *et al.*, Comparative genomic analysis of 45 type strains of the genus Bifidobacterium: a snapshot of its genetic diversity and evolution. *Plos One* **10**, e0117912 (2015).
30. V. Sharma, F. Mobeen, T. Prakash, Exploration of Survival Traits, Probiotic Determinants, Host Interactions, and Functional Evolution of Bifidobacterial Genomes Using Comparative Genomics. *Genes-Basel* **9**, 447 (2018).
31. C. I. Rodriguez, J. B. H. Martiny, Evolutionary relationships among bifidobacteria and their hosts and environments. *BMC Genomics* **21**, 26 (2020).
32. S. Koskiniemi, S. Sun, O. G. Berg, D. I. Andersson, Selection-driven gene loss in bacteria. *PLoS Genet.* **8**, e1002787 (2012).
33. C. Milani *et al.*, Genomics of the Genus Bifidobacterium Reveals Species-Specific Adaptation to the Glycan-Rich Gut Environment. *Appl. Environ. Microbiol.* **82**, 980-991 (2016).
34. S. Bentley, Sequencing the species pan-genome. *Nat. Rev. Microbiol.* **7**, 258-259 (2009).
35. P. Lapierre, J. P. Gogarten, Estimating the size of the bacterial pan-genome. *Trends Genet.* **25**, 107-110 (2009).
36. F. Bottacini *et al.*, Comparative genomics of the Bifidobacterium breve taxon. *BMC Genomics* **15**, 170 (2014).
37. B. Mayo, D. v. Sinderen, *Bifidobacteria : genomics and molecular aspects*. (Caister Academic, Norfolk, 2010), pp. 259.
38. C. Milani *et al.*, Bifidobacteria exhibit social behavior through carbohydrate resource sharing in the gut. *Sci Rep-Uk* **5**, 15782 (2015).
39. J. H. Lee, D. J. O'Sullivan, Genomic Insights into Bifidobacteria. *Microbiol. Mol. Biol. Rev.* **74**, 378 (2010).
40. M. Korakli, M. G. Ganzle, R. F. Vogel, Metabolism by bifidobacteria and lactic acid bacteria of polysaccharides from wheat and rye, and exopolysaccharides produced by Lactobacillus sanfranciscensis. *J. Appl. Microbiol.* **92**, 958-965 (2002).

41. L. V. Hooper, T. Midtvedt, J. I. Gordon, How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annu. Rev. Nutr.* **22**, 283-307 (2002).
42. K. Pokusaeva, G. F. Fitzgerald, D. van Sinderen, Carbohydrate metabolism in Bifidobacteria. *Genes Nutr* **6**, 285-306 (2011).
43. A. Walker, A. Cerdeno-Tarraga, S. Bentley, Faecal matters. *Nat. Rev. Microbiol.* **4**, 572-573 (2006).
44. M. A. Schell *et al.*, The genome sequence of Bifidobacterium longum reflects its adaptation to the human gastrointestinal tract. *P Natl Acad Sci USA* **99**, 14422-14427 (2002).
45. J. H. Lee *et al.*, Comparative genomic analysis of the gut bacterium Bifidobacterium longum reveals loci susceptible to deletion during pure culture growth. *BMC Genomics* **9**, 247 (2008).
46. J. F. Kim *et al.*, Genome sequence of the probiotic bacterium Bifidobacterium animalis subsp. lactis AD011. *J. Bacteriol.* **191**, 678-679 (2009).
47. P. W. Postma, J. W. Lengeler, G. R. Jacobson, Phosphoenolpyruvate:carbohydrate phosphotransferase systems of bacteria. *Microbiol Rev* **57**, 543-594 (1993).
48. A. O'Callaghan, D. van Sinderen, Bifidobacteria and Their Role as Members of the Human Gut Microbiota. *Front Microbiol* **7**, 925 (2016).
49. G. E. Felis, F. Dellaglio, Taxonomy of Lactobacilli and Bifidobacteria. *Curr Issues Intest Microbiol* **8**, 44-61 (2007).
50. R. J. Palframan, G. R. Gibson, R. A. Rastall, Carbohydrate preferences of Bifidobacterium species isolated from the human gut. *Curr Issues Intest Microbiol* **4**, 71-75 (2003).
51. G. Falony *et al.*, In vitro kinetic analysis of fermentation of prebiotic inulin-type fructans by Bifidobacterium species reveals four different phenotypes. *Appl. Environ. Microbiol.* **75**, 454-461 (2009).
52. E. N. Bergman, Energy contributions of volatile fatty acids from the gastrointestinal tract in various species. *Physiol. Rev.* **70**, 567-590 (1990).
53. G. D'Argenio, G. Mazzacca, Short-chain fatty acid in the human colon. Relation to inflammatory bowel diseases and colon cancer. *Adv. Exp. Med. Biol.* **472**, 149-158 (1999).
54. R. Correa-Oliveira, J. L. Fachi, A. Vieira, F. T. Sato, M. A. Vinolo, Regulation of immune cell function by short-chain fatty acids. *Clin Transl Immunology* **5**, e73 (2016).
55. A. Riviere, M. Selak, D. Lantin, F. Leroy, L. De Vuyst, Bifidobacteria and Butyrate-Producing Colon Bacteria: Importance and Strategies for Their Stimulation in the Human Gut. *Front Microbiol* **7**, 979 (2016).
56. M. Ventura *et al.*, Genome-scale analyses of health-promoting bacteria: probiogenomics. *Nat. Rev. Microbiol.* **7**, 61-71 (2009).
57. B. A. Degnan, G. T. Macfarlane, Transport and metabolism of glucose and arabinose in Bifidobacterium breve. *Arch. Microbiol.* **160**, 144-151 (1993).
58. A. El Kaoutari, F. Armougom, J. I. Gordon, D. Raoult, B. Henrissat, The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat. Rev. Microbiol.* **11**, 497-504 (2013).
59. D. A. Sela *et al.*, The genome sequence of Bifidobacterium longum subsp infantis reveals adaptations for milk utilization within the infant microbiome. *P Natl Acad Sci USA* **105**, 18964-18969 (2008).
60. E. Yoshida *et al.*, Bifidobacterium longum subsp. infantis uses two different beta-galactosidases for selectively degrading type-1 and type-2 human milk oligosaccharides. *Glycobiology* **22**, 361-368 (2012).

61. F. Turroni *et al.*, Bifidobacterium bifidum as an example of a specialized human gut commensal. *Front Microbiol* **5**, 437 (2014).
62. S. Duranti *et al.*, Insights from genomes of representatives of the human gut commensal Bifidobacterium bifidum. *Environ. Microbiol.* **17**, 2515-2531 (2015).
63. M. Kiyohara *et al.*, An exo-alpha-sialidase from bifidobacteria involved in the degradation of sialyloligosaccharides in human milk and intestinal glycoconjugates. *Glycobiology* **21**, 437-447 (2011).
64. M. Egan *et al.*, Cross-feeding by Bifidobacterium breve UCC2003 during co-cultivation with Bifidobacterium bifidum PRL2010 in a mucin-based medium. *BMC Microbiol.* **14**, 282 (2014).
65. S. C. Langley-Evans, Nutrition in early life and the programming of adult disease: a review. *J Hum Nutr Diet* **28 Suppl 1**, 1-14 (2015).
66. R. Nagpal *et al.*, Evolution of gut Bifidobacterium population in healthy Japanese infants over the first three years of life: a quantitative assessment. *Sci Rep* **7**, 10097 (2017).
67. E. G. Zoetendal, M. Rajilic-Stojanovic, W. M. de Vos, High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. *Gut* **57**, 1605-1615 (2008).
68. F. Turroni *et al.*, Microbiomic analysis of the bifidobacterial population in the human distal gut. *Isme J* **3**, 745-751 (2009).
69. G. W. Tannock *et al.*, Comparison of the compositions of the stool microbiotas of infants fed goat milk formula, cow milk-based formula, or breast milk. *Appl. Environ. Microbiol.* **79**, 3040-3048 (2013).
70. F. Backhed *et al.*, Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe* **17**, 852 (2015).
71. F. Sambo *et al.*, Optimizing PCR primers targeting the bacterial 16S ribosomal RNA gene. *BMC Bioinformatics* **19**, 343 (2018).
72. F. Turroni *et al.*, Diversity of Bifidobacteria within the Infant Gut Microbiota. *Plos One* **7**, e36957 (2012).
73. T. Matsuki *et al.*, Quantitative PCR with 16S rRNA-Gene-targeted species-specific primers for analysis of human intestinal bifidobacteria. *Appl. Environ. Microbiol.* **70**, 167-173 (2004).
74. A. Lucas, Programming by Early Nutrition in Man. *Ciba F Symp* **156**, 38-55 (1991).
75. M. Hanson, C. Fall, S. Robinson, J. Baird, *Early life nutrition and lifelong health*. (British Medical Association, London, 2009), pp. 122.
76. SACN, *Dietary Reference Values for Energy*. (Scientific Advisory Committee on Nutrition, London, 2011), pp. 228.
77. H. Renz, P. Brandtzaeg, M. Hornef, The impact of perinatal immune development on mucosal homeostasis and chronic inflammation. *Nature Reviews Immunology* **12**, 9-23 (2012).
78. P. J. Turnbaugh *et al.*, An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027-1031 (2006).
79. T. Olszak *et al.*, Microbial Exposure During Early Life Has Persistent Effects on Natural Killer T Cell Function. *Science* **336**, 489-493 (2012).
80. N. A. Bokulich *et al.*, Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science Translational Medicine* **8**, 343ra382 (2016).
81. R. Burcelin, Gut microbiota and immune crosstalk in metabolic disease. *Mol Metab* **5**, 771-781 (2016).
82. H. K. Pedersen *et al.*, Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* **535**, 376 (2016).

83. W. H. W. Tang, T. Kitai, S. L. Hazen, Gut Microbiota in Cardiovascular Health and Disease. *Circul. Res.* **120**, 1183-1196 (2017).
84. C. de Martel *et al.*, Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol* **13**, 607-615 (2012).
85. Q. Feng *et al.*, Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat Commun* **6**, 6528 (2015).
86. N. Voreades, A. Kozil, T. L. Weir, Diet and the development of the human intestinal microbiome. *Front Microbiol* **5**, 494 (2014).
87. R. K. Singh *et al.*, Influence of diet on the gut microbiome and implications for human health. *Journal of Translational Medicine* **15**, 73 (2017).
88. J. E. Koenig *et al.*, Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci U S A* **108 Suppl 1**, 4578-4585 (2011).
89. M. Fallani *et al.*, Determinants of the human infant intestinal microbiota after the introduction of first complementary foods in infant samples from five European centres. *Microbiology* **157**, 1385-1392 (2011).
90. M. F. Laursen *et al.*, Infant Gut Microbiota Development Is Driven by Transition to Family Foods Independent of Maternal Obesity. *Msphere* **1**, e00069-00015 (2016).
91. A. L. Thompson, A. Monteagudo-Mera, M. B. Cadenas, M. L. Lampl, M. A. Azcarate-Peril, Milk- and solid-feeding practices and daycare attendance are associated with differences in bacterial diversity, predominant communities, and metabolic and immune function of the infant gut microbiome. *Front Cell Infect Microbiol* **5**, 3 (2015).
92. R. Martin *et al.*, Early-Life Events, Including Mode of Delivery and Type of Feeding, Siblings and Gender, Shape the Developing Gut Microbiota. *Plos One* **11**, e0158498 (2016).
93. C. Milani *et al.*, The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. *Microbiol. Mol. Biol. Rev.* **81**, e00036-00017 (2017).
94. G. R. Gibson, M. B. Roberfroid, Dietary Modulation of the Human Colonic Microbiota - Introducing the Concept of Prebiotics. *J. Nutr.* **125**, 1401-1412 (1995).
95. M. E. Sanders, International Scientific Association for Probiotics and Prebiotics 2010 Meeting Report. *Functional Food Reviews* **2**, 131-140 (2010).
96. L. B. Bindels, N. M. Delzenne, P. D. Cani, J. Walter, Towards a more comprehensive concept for prebiotics. *Nat Rev Gastro Hepat* **12**, 303-310 (2015).
97. R. W. Hutkins *et al.*, Prebiotics: why definitions matter. *Curr. Opin. Biotechnol.* **37**, 1-7 (2016).
98. G. R. Gibson *et al.*, The International Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on the definition and scope of prebiotics. *Nat Rev Gastro Hepat* **14**, 491-502 (2017).
99. G. R. Gibson, Scott, K. P., Rastall, R. A., Tuohy, K. M., Hotchkiss, A., Dubert-Ferrandon, A., et al., Dietary prebiotics: current status and new definition. *Food Science & Technology Bulletin Functional Foods* **7(1)**, 1-19 (2010).
100. G. R. Gibson, H. M. Probert, J. Van Loo, R. A. Rastall, M. B. Roberfroid, Dietary modulation of the human colonic microbiota: updating the concept of prebiotics. *Nutr Res Rev* **17**, 259-275 (2004).
101. J. Stowell, Nutritional approaches to gut health: fibre, prebiotics, probiotics and synbiotics - A summary of definitions and implications of recent legislation. *Agro Food Ind Hi Tec* **18**, 20 (2007).
102. G. V. Coppa, S. Bruni, L. Morelli, S. Soldi, O. Gabrielli, The first prebiotics in humans - Human milk oligosaccharides. *J Clin Gastroenterol* **38**, S80-S83 (2004).

103. L. Bode, Human milk oligosaccharides: prebiotics and beyond. *Nutr. Rev.* **67**, S183-S191 (2009).
104. D. Barile, R. A. Rastall, Human milk and related oligosaccharides as prebiotics. *Curr. Opin. Biotechnol.* **24**, 214-219 (2013).
105. M. M. R. do Carmo *et al.*, Polydextrose: Physiological Function, and Effects on Health. *Nutrients* **8**, 553 (2016).
106. Y. Wu *et al.*, Effects of isomalto-oligosaccharides as potential prebiotics on performance, immune function and gut microbiota in weaned pigs. *Anim. Feed Sci. Technol.* **230**, 126-135 (2017).
107. P. Markowiak, K. Slizewska, Effects of Probiotics, Prebiotics, and Synbiotics on Human Health. *Nutrients* **9**, 1021 (2017).
108. S. L. Collins *et al.*, Promising Prebiotic Candidate Established by Evaluation of Lactitol, Lactulose, Raffinose, and Oligofructose for Maintenance of a Lactobacillus-Dominated Vaginal Microbiota. *Appl. Environ. Microbiol.* **84**, e02200-02217 (2018).
109. M. Sabater-Molina, E. Larque, F. Torrella, S. Zamora, Dietary fructooligosaccharides and potential benefits on health. *J Physiol Biochem* **65**, 315-328 (2009).
110. B. H. Xu, Y. B. Wang, J. R. Li, Q. Lin, Effect of prebiotic xylooligosaccharides on growth performances and digestive enzyme activities of allogynogenetic crucian carp (Carassius auratus gibelio). *Fish Physiol. Biochem.* **35**, 351-357 (2009).
111. V. Mandal, S. K. Sen, N. C. Mandal, Effect of prebiotics on bacteriocin production and cholesterol lowering activity of Pediococcus acidilactici LAB 5. *World J. Microbiol. Biotechnol.* **25**, 1837-1847 (2009).
112. E. Vamanu, A. Vamanu, The influence of prebiotics on bacteriocin synthesis using the strain Lactobacillus paracasei CMGB16. *Afr J Microbiol Res* **4**, 534-537 (2010).
113. S. K. Yeo, M. T. Liong, Effect of prebiotics on viability and growth characteristics of probiotics in soymilk. *J. Sci. Food Agric.* **90**, 267-275 (2010).
114. E. Fuentes-Zaragoza *et al.*, Resistant starch as prebiotic: A review. *Starch-Starke* **63**, 406-415 (2011).
115. R. H. Vaidya, M. K. Sheth, Processing and storage of Indian cereal and cereal products alters its resistant starch content. *J Food Sci Tech Mys* **48**, 622-627 (2011).
116. J. Carlson, J. Slavin, Health benefits of fibre, prebiotics and probiotics: a review of intestinal health and related health claims. *Qual Assur Saf Crop* **8**, 539-553 (2016).
117. O. Ballard, A. L. Morrow, Human milk composition: nutrients and bioactive factors. *Pediatr Clin North Am* **60**, 49-74 (2013).
118. D. Garwolinska, J. Namiesnik, A. Kot-Wasik, W. Hewelt-Belka, Chemistry of Human Breast Milk-A Comprehensive Review of the Composition and Role of Milk Metabolites in Child Development. *J. Agric. Food Chem.* **66**, 11881-11896 (2018).
119. I. Le Huerou-Luron, S. Blat, G. Boudry, Breast- v. formula-feeding: impacts on the digestive tract and immediate and long-term health effects. *Nutr Res Rev* **23**, 23-36 (2010).
120. C. Gomez-Gallego, I. Garcia-Mantrana, S. Salminen, M. C. Collado, The human milk microbiome and factors influencing its composition and activity. *Seminars in fetal & neonatal medicine* **21**, 400-405 (2016).
121. V. Vieira Borba, K. Sharif, Y. Shoenfeld, Breastfeeding and autoimmunity: Programing health from the beginning. *Am. J. Reprod. Immunol.* **79**, e12778 (2018).
122. T. Hennet, L. Borsig, Breastfed at Tiffany's. *Trends Biochem. Sci.* **41**, 508-518 (2016).
123. A. Boix-Amoros *et al.*, Reviewing the evidence on breast milk composition and immunological outcomes. *Nutr. Rev.* **77**, 541-556 (2019).
124. D. Munblit *et al.*, Immune Components in Human Milk Are Associated with Early Infant Immunological Health Outcomes: A Prospective Three-Country Analysis. *Nutrients* **9**, 532 (2017).

125. A. Stuebe, The risks of not breastfeeding for mothers and infants. *Rev Obstet Gynecol* **2**, 222-231 (2009).
126. M. L. Forchielli, W. A. Walker, The role of gut-associated lymphoid tissues and mucosal defence. *Br J Nutr* **93 Suppl 1**, S41-48 (2005).
127. J. A. Peterson, S. Patton, M. Hamosh, Glycoproteins of the human milk fat globule in the protection of the breast-fed infant against infections. *Biol Neonate* **74**, 143-162 (1998).
128. B. Andersson, O. Porras, L. A. Hanson, T. Lagergard, C. Svanborg-Eden, Inhibition of attachment of Streptococcus pneumoniae and Haemophilus influenzae by human milk and receptor oligosaccharides. *J. Infect. Dis.* **153**, 232-237 (1986).
129. B. Andersson *et al.*, Identification of an active disaccharide unit of a glycoconjugate receptor for pneumococci attaching to human pharyngeal epithelial cells. *J. Exp. Med.* **158**, 559-570 (1983).
130. S. Lehmann *et al.*, In Vitro Evidence for Immune-Modulatory Properties of Non-Digestible Oligosaccharides: Direct Effect on Human Monocyte Derived Dendritic Cells. *Plos One* **10**, e0132304 (2015).
131. M. A. Naarding *et al.*, Lewis X component in human milk binds DC-SIGN and inhibits HIV-1 transfer to CD4+ T lymphocytes. *J. Clin. Invest.* **115**, 3256-3264 (2005).
132. M. P. Heikkila, P. E. Saris, Inhibition of Staphylococcus aureus by the commensal bacteria of human milk. *J. Appl. Microbiol.* **95**, 471-478 (2003).
133. R. Martin *et al.*, Isolation of bifidobacteria from breast milk and assessment of the bifidobacterial population by PCR-denaturing gradient gel electrophoresis and quantitative real-time PCR. *Appl. Environ. Microbiol.* **75**, 965-969 (2009).
134. T. Jost, C. Lacroix, C. Braegger, C. Chassard, Assessment of bacterial diversity in breast milk using culture-dependent and culture-independent approaches. *Br J Nutr* **110**, 1253-1262 (2013).
135. M. Gueimonde, K. Laitinen, S. Salminen, E. Isolauri, Breast milk: a source of bifidobacteria for infant gut development and maturation? *Neonatology* **92**, 64-66 (2007).
136. K. M. Hunt *et al.*, Characterization of the diversity and temporal stability of bacterial communities in human milk. *Plos One* **6**, e21313 (2011).
137. R. Cabrera-Rubio *et al.*, The human milk microbiome changes over lactation and is shaped by maternal weight and mode of delivery. *Am. J. Clin. Nutr.* **96**, 544-551 (2012).
138. S. Moossavi *et al.*, Composition and Variation of the Human Milk Microbiota Are Influenced by Maternal and Early-Life Factors. *Cell Host Microbe* **25**, 324-335 (2019).
139. K. A. Lackey *et al.*, What's Normal? Microbiomes in Human Milk and Infant Feces Are Related to Each Other but Vary Geographically: The INSPIRE Study. *Front Nutr* **6**, 45 (2019).
140. L. Ruiz, C. Garcia-Carral, J. M. Rodriguez, Unfolding the Human Milk Microbiome Landscape in the Omics Era. *Front Microbiol* **10**, 1378 (2019).
141. S. H. Patel *et al.*, Culture independent assessment of human milk microbial community in lactational mastitis. *Sci Rep* **7**, 7804 (2017).
142. S. Duranti *et al.*, Maternal inheritance of bifidobacterial communities and bifidophages in infants through vertical transmission. *Microbiome* **5**, 66 (2017).
143. M. Yassour *et al.*, Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. *Cell Host Microbe* **24**, 146-154 (2018).
144. T. Jost, C. Lacroix, C. P. Braegger, F. Rochat, C. Chassard, Vertical mother-neonate transfer of maternal gut bacteria via breastfeeding. *Environ. Microbiol.* **16**, 2891-2904 (2014).

145. E. Biagi *et al.*, The Bacterial Ecosystem of Mother's Milk and Infant's Mouth and Gut. *Front Microbiol* **8**, 1214 (2017).
146. H. Makino *et al.*, Mother-to-infant transmission of intestinal bifidobacterial strains has an impact on the early development of vaginally delivered infant's microbiota. *PLoS One* **8**, e78331 (2013).
147. H. Makino *et al.*, Transmission of intestinal Bifidobacterium longum subsp. longum strains from mother to infant, determined by multilocus sequencing typing and amplified fragment length polymorphism. *Appl. Environ. Microbiol.* **77**, 6788-6793 (2011).
148. D. Benito *et al.*, Characterization of Staphylococcus aureus strains isolated from faeces of healthy neonates and potential mother-to-infant microbial transmission through breastfeeding. *FEMS Microbiol. Ecol.* **91**, fiv007 (2015).
149. M. Dzidic, A. Boix-Amoros, M. Selma-Royo, A. Mira, M. C. Collado, Gut Microbiota and Mucosal Immunity in the Neonate. *Med Sci* **6**, 56 (2018).
150. L. Fernandez *et al.*, The human milk microbiota: origin and potential roles in health and disease. *Pharmacol. Res.* **69**, 1-10 (2013).
151. D. Sprockett, T. Fukami, D. A. Relman, Role of priority effects in the early-life assembly of the gut microbiota. *Nature reviews. Gastroenterology & hepatology* **15**, 197-205 (2018).
152. A. Gotoh *et al.*, Sharing of human milk oligosaccharides degradants within bifidobacterial communities in faecal cultures supplemented with Bifidobacterium bifidum. *Sci Rep-Uk* **8**, 13958 (2018).
153. A. Marcobal, J. L. Sonnenburg, Human milk oligosaccharide consumption by intestinal microbiota. *Clin. Microbiol. Infect.* **18**, 12-15 (2012).
154. M. R. Ninonuevo *et al.*, Daily variations in oligosaccharides of human milk determined by microfluidic chips and mass spectrometry. *J. Agric. Food Chem.* **56**, 618-626 (2008).
155. C. Yamada *et al.*, Molecular Insight into Evolution of Symbiosis between Breast-Fed Infants and a Member of the Human Gut Microbiome Bifidobacterium longum. *Cell Chem Biol* **24**, 515 (2017).
156. K. James, M. O. Motherway, F. Bottacini, D. van Sinderen, Bifidobacterium breve UCC2003 metabolises the human milk oligosaccharides lacto-N-tetraose and lacto-N-neo-tetraose through overlapping, yet distinct pathways. *Sci Rep* **6**, 38560 (2016).
157. T. Katayama, Host-derived glycans serve as selected nutrients for the gut microbe: human milk oligosaccharides and bifidobacteria. *Biosci Biotech Bioch* **80**, 621-632 (2016).
158. D. Garrido *et al.*, A novel gene cluster allows preferential utilization of fucosylated milk oligosaccharides in Bifidobacterium longum subsp longum SC596. *Sci Rep-Uk* **6**, 35045 (2016).
159. P. Thomson, D. A. Medina, D. Garrido, Human milk oligosaccharides and infant gut bifidobacteria: Molecular strategies for their utilization. *Food Microbiol.* **75**, 37-46 (2018).
160. T. Odamaki *et al.*, Comparative Genomics Revealed Genetic Diversity and Species/Strain-Level Differences in Carbohydrate Metabolism of Three Probiotic Bifidobacterial Species. *Int J Genomics* **2015**, 567809 (2015).
161. S. Asakuma *et al.*, Physiology of Consumption of Human Milk Oligosaccharides by Infant Gut-associated Bifidobacteria. *J. Biol. Chem.* **286**, 34583-34592 (2011).
162. M. A. E. Lawson *et al.*, Breast milk-derived human milk oligosaccharides promote Bifidobacterium interactions within a single ecosystem. *Isme J* **14**, 635-648 (2020).

163. Z. T. Yu, C. Chen, D. S. Newburg, Utilization of major fucosylated and sialylated human milk oligosaccharides by isolated human gut microbes. *Glycobiology* **23**, 1281-1292 (2013).
164. D. Garrido *et al.*, Utilization of galactooligosaccharides by Bifidobacterium longum subsp. infantis isolates. *Food Microbiol.* **33**, 262-270 (2013).
165. I. S. O'Neill, Z.; Hall, Lindsay J, Exploring the role of the microbiota member Bifidobacterium in modulating immune-linked diseases. *Emerging Topics in Life Sciences* **1**, 333-349 (2017).
166. L. Vitetta, G. Vitetta, S. Hall, Immunological Tolerance and Function: Associations Between Intestinal Bacteria, Probiotics, Prebiotics, and Phages. *Frontiers in Immunology* **9**, 2240 (2018).
167. M. Luu *et al.*, Regulation of the effector function of CD8(+) T cells by gut microbiota-derived metabolite butyrate. *Sci Rep* **8**, 14430 (2018).
168. M. Li *et al.*, Pro- and anti-inflammatory effects of short chain fatty acids on immune and endothelial cells. *Eur. J. Pharmacol.* **831**, 52-59 (2018).
169. R. M. Kent, S. B. Doherty, Probiotic bacteria in infant formula and follow-up formula: Microencapsulation using milk and pea proteins to improve microbiological quality. *Food Res. Int.* **64**, 567-576 (2014).
170. H. Yoshioka, K. Iseki, K. Fujita, Development and differences of intestinal flora in the neonatal period in breast-fed and bottle-fed infants. *Pediatrics* **72**, 317-321 (1983).
171. H. J. Harmsen *et al.*, Analysis of intestinal flora development in breast-fed and formula-fed infants by using molecular identification and detection methods. *J Pediatr Gastroenterol Nutr* **30**, 61-67 (2000).
172. A. O'Sullivan, M. Farver, J. T. Smilowitz, The Influence of Early Infant-Feeding Practices on the Intestinal Microbiome and Body Composition in Infants. *Nutr Metab Insights* **8**, 1-9 (2015).
173. C. Y. Lu, Y. H. Ni, Gut microbiota and the development of pediatric diseases. *J Gastroenterol* **50**, 720-726 (2015).
174. L. T. Stiemsma, K. B. Michels, The Role of the Microbiome in the Developmental Origins of Health and Disease. *Pediatrics* **141**, e20172437 (2018).
175. E. Bezirtzoglou, A. Tsiotsias, G. W. Welling, Microbiota profile in feces of breast- and formula-fed newborns by using fluorescence in situ hybridization (FISH). *Anaerobe* **17**, 478-482 (2011).
176. C. J. Stewart *et al.*, Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583 (2018).
177. T. Harder, R. Bergmann, G. Kallischnigg, A. Plagemann, Duration of breastfeeding and risk of overweight: a meta-analysis. *Am. J. Epidemiol.* **162**, 397-403 (2005).
178. S. Ip *et al.*, Breastfeeding and maternal and infant health outcomes in developed countries. *Evid Rep Technol Assess (Full Rep)*, 1-186 (2007).
179. U. N. Das, Breastfeeding prevents type 2 diabetes mellitus: but, how and why? *Am. J. Clin. Nutr.* **85**, 1436-1437 (2007).
180. J. A. Ortega-Garcia *et al.*, Full Breastfeeding and Obesity in Children: A Prospective Study from Birth to 6 Years. *Child Obes* **14**, 327-337 (2018).
181. R. M. Martin, D. Gunnell, G. D. Smith, Breastfeeding in infancy and blood pressure in later life: Systematic review and meta-analysis. *Am. J. Epidemiol.* **161**, 15-26 (2005).
182. C. G. Owen, P. H. Whincup, K. Odoki, J. A. Gilg, D. G. Cook, Infant feeding and blood cholesterol: A study in adolescents and a systematic review. *Pediatrics* **110**, 597-608 (2002).

183. K. E. Mercer *et al.*, Infant Formula Feeding Increases Hepatic Cholesterol 7 alpha Hydroxylase (CYP7A1) Expression and Fecal Bile Acid Loss in Neonatal Piglets. *J. Nutr.* **148**, 702-711 (2018).
184. C. R. Martin, P. R. Ling, G. L. Blackburn, Review of Infant Feeding: Key Features of Breast Milk and Infant Formula. *Nutrients* **8**, 279 (2016).
185. B. Koletzko *et al.*, Global standard for the composition of infant formula: Recommendations of an ESPGHAN Coordinated International Expert Group. *J Pediatr Gastr Nutr* **41**, 584-599 (2005).
186. K. Papagaroufalis, A. Fotiou, D. Egli, L. A. Tran, P. Steenhout, A Randomized Double Blind Controlled Safety Trial Evaluating d-Lactic Acid Production in Healthy Infants Fed a Lactobacillus reuteri-containing Formula. *Nutr Metab Insights* **7**, 19-27 (2014).
187. D. A. Cook, Nutrient levels in infant formulas: technical considerations. *J. Nutr.* **119**, 1773-1778 (1989).
188. C. Kunz, S. Rudloff, W. Baier, N. Klein, S. Strobel, Oligosaccharides in human milk: Structural, functional, and metabolic aspects. *Annu. Rev. Nutr.* **20**, 699-722 (2000).
189. Y. Vandenplas, I. Zakharova, Y. Dmitrieva, Oligosaccharides in infant formula: more evidence to validate the role of prebiotics. *Brit J Nutr* **113**, 1339-1344 (2015).
190. D. L. Ackerman, K. M. Craft, S. D. Townsend, Infant food applications of complex carbohydrates: Structure, synthesis, and function. *Carbohyd Res* **437**, 16-27 (2017).
191. S. Verkhnyatskaya, M. Ferrari, P. de Vos, M. T. C. Walvoort, Shaping the Infant Microbiome With Non-digestible Carbohydrates. *Front Microbiol* **10**, 343 (2019).
192. X. M. Ben *et al.*, Supplementation of milk formula with galacto-oligosaccharides improves intestinal micro-flora and fermentation in term infants. *Chinese Med J-Peking* **117**, 927-931 (2004).
193. S. Fanaro *et al.*, Galacto-oligosaccharides and long-chain fructo-oligosaccharides as prebiotics in infant formulas: A review. *Acta Paediatr* **94**, 22-26 (2005).
194. A. J. H. Maathuis, E. C. van den Heuvel, M. H. C. Schoterman, K. Venema, Galacto-Oligosaccharides Have Prebiotic Activity in a Dynamic In Vitro Colon Model Using a C-13-Labeling Technique. *J. Nutr.* **142**, 1205-1212 (2012).
195. D. Watson *et al.*, Selective carbohydrate utilization by lactobacilli and bifidobacteria. *J. Appl. Microbiol.* **114**, 1132-1146 (2013).
196. C. Sierra *et al.*, Prebiotic effect during the first year of life in healthy infants fed formula containing GOS as the only prebiotic: a multicentre, randomised, double-blind and placebo-controlled trial. *Eur J Nutr* **54**, 89-99 (2015).
197. S. P. Marx, S. Winkler, W. Hartmeier, Metabolization of beta-(2,6)-linked fructose-oligosaccharides by different bifidobacteria. *FEMS Microbiol. Lett.* **182**, 163-169 (2000).
198. M. Rossi *et al.*, Fermentation of fructooligosaccharides and inulin by bifidobacteria: a comparative study of pure and fecal cultures. *Appl. Environ. Microbiol.* **71**, 6150-6158 (2005).
199. H. Kaplan, R. W. Hutkins, Fermentation of fructooligosaccharides by lactic acid bacteria and bifidobacteria. *Appl. Environ. Microbiol.* **66**, 2682-2684 (2000).
200. R. K. Buddington, C. H. Williams, S. C. Chen, S. A. Witherly, Dietary supplement of neosugar alters the fecal flora and decreases activities of some reductive enzymes in human subjects. *Am. J. Clin. Nutr.* **63**, 709-716 (1996).
201. Y. Guigoz, F. Rochat, G. Perruisseau-Carrier, I. Rochat, E. J. Schiffrin, Effects of oligosaccharide on the faecal flora and non-specific immune system in elderly people. *Nutr Res* **22**, 13-25 (2002).

202. Y. Bouhnik *et al.*, Four-week short chain fructo-oligosaccharides ingestion leads to increasing fecal bifidobacteria and cholesterol excretion in healthy elderly volunteers. *Nutr J* **6**, 42 (2007).
203. C. Costalos, A. Kapiki, M. Apostolou, E. Papathoma, The effect of a prebiotic supplemented formula on growth and stool microbiology of term infants. *Early Hum Dev* **84**, 45-49 (2008).
204. C. A. Cherrington, M. Hinton, G. R. Pearson, I. Chopra, Short-Chain Organic-Acids at Ph 5.0 Kill Escherichia-Coli and Salmonella Spp without Causing Membrane Perturbation. *Journal of Applied Bacteriology* **70**, 161-165 (1991).
205. S. H. Duncan, P. Louis, J. M. Thomson, H. J. Flint, The role of pH in determining the species composition of the human colonic microbiota. *Environ. Microbiol.* **11**, 2112-2122 (2009).
206. D. Paineau *et al.*, Effects of Short-Chain Fructooligosaccharides on Faecal Bifidobacteria and Specific Immune Response in Formula-Fed Term Infants: A Randomized, Double-Blind, Placebo-Controlled Trial. *J. Nutr. Sci. Vitaminol.* **60**, 167-175 (2014).
207. D. Mohnen, Pectin structure and biosynthesis. *Curr. Opin. Plant Biol.* **11**, 266-277 (2008).
208. R. Di *et al.*, Pectic oligosaccharide structure-function relationships: Prebiotics, inhibitors of Escherichia coli O157:H7 adhesion and reduction of Shiga toxin cytotoxicity in HT29 cells. *Food Chem.* **227**, 245-254 (2017).
209. M. Lopez-Siles *et al.*, Cultured Representatives of Two Major Phylogroups of Human Colonic Faecalibacterium prausnitzii Can Utilize Pectin, Uronic Acids, and Host-Derived Substrates for Growth. *Appl. Environ. Microbiol.* **78**, 420-428 (2012).
210. P. Louis, H. J. Flint, C. Michel, How to Manipulate the Microbiota: Prebiotics. *Microbiota of the Human Body: Implications in Health and Disease* **902**, 119-142 (2016).
211. J. Stam, M. van Stuijvenberg, J. Garssen, K. Knipping, P. J. Sauer, A mixture of three prebiotics does not affect vaccine specific antibody responses in healthy term infants in the first year of life. *Vaccine* **29**, 7766-7772 (2011).
212. A. P. Vos *et al.*, Dietary supplementation of neutral and acidic oligosaccharides enhances Th1-dependent vaccination responses in mice. *Pediatr Allergy Immunol* **18**, 304-312 (2007).
213. A. P. Vos *et al.*, Dietary supplementation with specific oligosaccharide mixtures decreases parameters of allergic asthma in mice. *Int. Immunopharmacol.* **7**, 1582-1587 (2007).
214. D. Bosscher, M. Van Caillie-Bertrand, R. Van Cauwenbergh, H. Deelstra, Availabilities of calcium, iron, and zinc from dairy infant formulas is affected by soluble dietary fibers and modified starch fractions. *Nutrition* **19**, 641-645 (2003).
215. C. C. Chen *et al.*, Sequential one-pot multienzyme (OPME) synthesis of lacto-N-neotetraose and its sialyl and fucosyl derivatives. *Chem Commun* **51**, 7689-7692 (2015).
216. H. Yu *et al.*, H-pylori alpha 1-3/4-fucosyltransferase (Hp3/4FT)-catalyzed one-pot multienzyme (OPME) synthesis of Lewis antigens and human milk fucosides. *Chem Commun* **53**, 11012-11015 (2017).
217. H. M. Qin *et al.*, Multienzymatic cascade synthesis of fucosyloligosaccharide via a two-step fermentation strategy in Escherichia coli. *Biotechnol. Lett.* **38**, 1747-1752 (2016).
218. Y. W. Chin, J. Y. Kim, J. H. Kim, S. M. Jung, J. H. Seo, Improved production of 2 '-fucosyllactose in engineered Escherichia coli by expressing putative alpha-1,2-

fucosyltransferase, Wcf from Bacteroides fragilis. *J. Biotechnol.* **257**, 192-198 (2017).
219. S. Yu *et al.*, Production of a human milk oligosaccharide 2 '-fucosyllactose by metabolically engineered Saccharomyces cerevisiae. *Microb Cell Fact* **17**, 101 (2018).
220. P. M. Danby, S. G. Withers, Advances in Enzymatic Glycoside Synthesis. *Acs Chem Biol* **11**, 1784-1794 (2016).
221. N. H. A. Manas, R. M. Illias, N. M. Mahadi, Strategy in manipulating transglycosylation activity of glycosyl hydrolase for oligosaccharide production. *Crit. Rev. Biotechnol.* **38**, 272-293 (2018).
222. B. Zeuner, C. Jers, J. D. Mikkelsen, A. S. Meyer, Methods for Improving Enzymatic Trans-glycosylation for Synthesis of Human Milk Oligosaccharide Biomimetics. *J. Agric. Food Chem.* **62**, 9615-9631 (2014).
223. EC, Commission Implementing Decision (EU) 2016/376 of 11 March 2016 authorising the placing on the market of 2'-O-fucosyllactose as a novel food ingredient under Regulation (EC) No 258/97 of the European Parliament and of the Council (notified under document C(2016) 1423). Available at: https://eur-lex.europa.eu/eli/dec_impl/2016/2376/oj (2016).
224. EC, Commission Implementing Decision (EU) 2016/375 of 11 March 2016 authorising the placing on the market of lacto-N-neotetraose as a novel food ingredient under Regulation (EC) No 258/97 of the European Parliament and of the Council (notified under document C(2016) 1419). Available at: https://eur-lex.europa.eu/eli/dec_impl/2016/2375/oj (2016).
225. EC, Commission Implementing Decision (EU) 2017/2375 of 15 December 2017 authorising the placing on the market of N-acetyl-D-neuraminic acid as a novel food ingredient under Regulation (EC) No 258/97 of the European Parliament and of the Council (notified under document C(2017) 8431). Available at: https://eur-lex.europa.eu/eli/dec_impl/2017/2375/oj (2017).
226. EC, Commission Implementing Decision (EU) 2017/2201 of 27 November 2017 authorising the placing on the market of 2'-fucosyllactose produced with Escherichia coli strain BL21 as a novel food ingredient under Regulation (EC) No 258/97 of the European Parliament and of the Council (notified under document C(2017) 7662). Available at: https://eur-lex.europa.eu/eli/dec_impl/2017/2201/oj (2017).
227. WHO/FAO, Report of a joint FAO/WHO expert consultation on evaluation of health and nutritional properties of probiotics in food including powder milk with live lactic acid bacteria. Available at: http://www.fao.org/publications/card/en/c/7c102d195-102fd105-105b122-108faf-f100b102e168dfbb106/ (2001).
228. C. Hill *et al.*, The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic. *Nat Rev Gastro Hepat* **11**, 506-514 (2014).
229. R. Fuller, Probiotics in Man and Animals. *Journal of Applied Bacteriology* **66**, 365-378 (1989).
230. J. H. J. H. Veld, R. Havenaar, Probiotics and Health in Man and Animal. *J. Chem. Technol. Biotechnol.* **51**, 562-567 (1991).
231. A. K. Anal, H. Singh, Recent advances in microencapsulation of probiotics for industrial applications and targeted delivery. *Trends Food Sci. Technol.* **18**, 240-251 (2007).

232. P. de Vos, M. M. Faas, M. Spasojevic, J. Sikkema, Encapsulation for preservation of functionality and targeted delivery of bioactive food components. *Int. Dairy J.* **20**, 292-302 (2010).
233. M. N. Mugambi, A. Musekiwa, M. Lombard, T. Young, R. Blaauw, Synbiotics, probiotics or prebiotics in infant formula for full term infants: a systematic review. *Nutr J* **11**, 81 (2012).
234. C. Braegger *et al.*, Supplementation of Infant Formula With Probiotics and/or Prebiotics: A Systematic Review and Comment by the ESPGHAN Committee on Nutrition. *J Pediatr Gastr Nutr* **52**, 238-250 (2011).
235. EFSA, Introduction of a Qualified Presumption of Safety (QPS) approach for assessment of selected microorganisms referred to EFSA - Opinion of the Scientific Committee. *EFSA Journal*, 587 (2007).
236. FDA, Evidence-based review system for the scientific evaluation of health claim. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-industry-evidence-based-review-system-scientific-evaluation-health-claims (2009).
237. EC, Regulation (EC) No 1924/2006 of the European Parliament and of the Council of 20 December 2006 on nutrition and health claims made on foods. Available at: https://eur-lex.europa.eu/eli/reg/2006/1924/oj (2006).
238. M. E. Sanders, D. Merenstein, C. A. Merrifield, R. Hutkins, Probiotics for human use. *Nutr Bull* **43**, 212-225 (2018).
239. M. E. Sanders *et al.*, Safety assessment of probiotics for human use. *Gut Microbes* **1**, 164-185 (2010).
240. M. van den Nieuwboer, E. Claassen, L. Morelli, F. Guarner, R. J. Brummer, Probiotic and synbiotic safety in infants under two years of age. *Beneficial Microbes* **5**, 45-60 (2014).
241. S. McKeen *et al.*, Infant Complementary Feeding of Prebiotics for the Microbiome and Immunity. *Nutrients* **11**, 364 (2019).
242. S. Hald *et al.*, Effects of Arabinoxylan and Resistant Starch on Intestinal Microbiota and Short-Chain Fatty Acids in Subjects with Metabolic Syndrome: A Randomised Crossover Study. *Plos One* **11**, e0159223 (2016).
243. M. B. Roberfroid, Inulin-type fructans: functional food ingredients. *J. Nutr.* **137**, 2493S-2502S (2007).
244. W. F. Broekaert *et al.*, Prebiotic and other health-related effects of cereal-derived arabinoxylans, arabinoxylan-oligosaccharides, and xylooligosaccharides. *Crit. Rev. Food Sci. Nutr.* **51**, 178-194 (2011).
245. K. R. Niness, Inulin and oligofructose: What are they? *J. Nutr.* **129**, 1402s-1406s (1999).
246. K. P. Scott *et al.*, Substrate-driven gene expression in Roseburia inulinivorans: Importance of inducible enzymes in the utilization of inulin and starch. *P Natl Acad Sci USA* **108**, 4672-4679 (2011).
247. M. Warchol, S. Perrin, J. P. Grill, F. Schneider, Characterization of a purified beta-fructofuranosidase from Bifidobacterium infantis ATCC 15697. *Lett. Appl. Microbiol.* **35**, 462-467 (2002).
248. M. A. Ehrmann, M. Korakli, R. F. Vogel, Identification of the gene for beta-fructofuranosidase of Bifidobacterium lactis DSM10140(T) and characterization of the enzyme expressed in Escherichia coli. *Curr. Microbiol.* **46**, 391-397 (2003).
249. T. Omori *et al.*, Characterization of recombinant beta-fructofuranosidase from Bifidobacterium adolescentis G1. *Chem Cent J* **4**, 9 (2010).

250. G. Falony *et al.*, In Vitro Kinetic Analysis of Fermentation of Prebiotic Inulin-Type Fructans by Bifidobacterium Species Reveals Four Different Phenotypes. *Appl. Environ. Microbiol.* **75**, 454-461 (2009).
251. M. Selak *et al.*, Inulin-type fructan fermentation by bifidobacteria depends on the strain rather than the species and region in the human intestine. *Appl. Microbiol. Biotechnol.* **100**, 4097-4107 (2016).
252. G. Falony, A. Vlachou, K. Verbrugghe, L. De Vuyst, Cross-feeding between Bifidobacterium longum BB536 and acetate-converting, butyrate-producing colon bacteria during growth on oligofructose. *Appl. Environ. Microbiol.* **72**, 7835-7841 (2006).
253. G. Falony *et al.*, In Vitro Kinetics of Prebiotic Inulin-Type Fructan Fermentation by Butyrate-Producing Colon Bacteria: Implementation of Online Gas Chromatography for Quantitative Analysis of Carbon Dioxide and Hydrogen Gas Production. *Appl. Environ. Microbiol.* **75**, 5884-5892 (2009).
254. F. Moens, S. Weckx, L. De Vuyst, Bifidobacterial inulin-type fructan degradation capacity determines cross-feeding interactions between bifidobacteria and Faecalibacterium prausnitzii. *Int. J. Food Microbiol.* **231**, 76-85 (2016).
255. G. Schaafsma, J. L. Slavin, Significance of Inulin Fructans in the Human Diet. *Compr Rev Food Sci F* **14**, 37-47 (2015).
256. M. S. Izydorczyk, C. G. Biliaderis, Cereal arabinoxylans: Advances in structure and physicochemical properties. *Carbohyd Polym* **28**, 33-48 (1995).
257. N. K. Morgan, C. Keerqin, A. Wallace, S. B. Wu, M. Choct, Effect of arabinoxylo-oligosaccharides and arabinoxylans on net energy and nutrient utilization in broilers. *Anim Nutr* **5**, 56-62 (2019).
258. L. A. M. Van Den Broek, A. G. J. Voragen, Bifidobacterium glycoside hydrolases and (potential) prebiotics. *Innov Food Sci Emerg* **9**, 401-407 (2008).
259. S. Lagaert, A. Pollet, C. M. Courtin, G. Volckaert, beta-Xylosidases and alpha-L-arabinofuranosidases: Accessory enzymes for arabinoxylan degradation. *Biotechnol. Adv.* **32**, 316-332 (2014).
260. S. Lagaert *et al.*, Characterization of two beta-xylosidases from Bifidobacterium adolescentis and their contribution to the hydrolysis of prebiotic xylooligosaccharides. *Appl. Microbiol. Biotechnol.* **92**, 1179-1185 (2011).
261. S. Lagaert *et al.*, Substrate specificity of three recombinant alpha-L-arabinofuranosidases from Bifidobacterium adolescentis and their divergent action on arabinoxylan and arabinoxylan oligosaccharides. *Biochem. Biophys. Res. Commun.* **402**, 644-650 (2010).
262. S. A. Hughes *et al.*, In vitro fermentation by human fecal microflora of wheat arabinoxylans. *J. Agric. Food Chem.* **55**, 4589-4595 (2007).
263. W. S. F. Chung *et al.*, Relative abundance of the Prevotella genus within the human gut microbiota of elderly volunteers determines the inter-individual responses to dietary supplementation with wheat bran arabinoxylan-oligosaccharides. *BMC Microbiol.* **20**, 283 (2020).
264. V. Van Craeyveld *et al.*, Structurally Different Wheat-Derived Arabinoxylooligosaccharides Have Different Prebiotic and Fermentation Properties in Rats. *J. Nutr.* **138**, 2348-2355 (2008).
265. P. Van den Abbeele *et al.*, Arabinoxylans and inulin differentially modulate the mucosal and luminal gut microbiota and mucin-degradation in humanized rats. *Environ. Microbiol.* **13**, 2667-2680 (2011).
266. P. Truchado *et al.*, Bifidobacterium longum D2 enhances microbial degradation of long-chain arabinoxylans in an in vitro model of the proximal colon. *Beneficial Microbes* **6**, 849-860 (2015).

267. A. Riviere *et al.*, The Ability of Bifidobacteria To Degrade Arabinoxylan Oligosaccharide Constituents and Derived Oligosaccharides Is Strain Dependent. *Appl. Environ. Microbiol.* **80**, 204-217 (2014).
268. C. Grootaert *et al.*, Comparison of prebiotic effects of arabinoxylan oligosaccharides and inulin in a simulator of the human intestinal microbial ecosystem. *FEMS Microbiol. Ecol.* **69**, 231-242 (2009).
269. J. I. Sanchez *et al.*, Arabinoxylan-oligosaccharides (AXOS) affect the protein/carbohydrate fermentation balance and microbial population dynamics of the Simulator of Human Intestinal Microbial Ecosystem. *Microb Biotechnol* **2**, 101-113 (2009).
270. A. M. Neyrinck *et al.*, Wheat-derived arabinoxylan oligosaccharides with prebiotic effect increase satietogenic gut peptides and reduce metabolic endotoxemia in diet-induced obese mice. *Nutr Diabetes* **2**, e28 (2012).
271. C. Grootaert *et al.*, Microbial metabolism and prebiotic potency of arabinoxylan oligosaccharides in the human intestine. *Trends Food Sci. Technol.* **18**, 64-71 (2007).
272. B. Damen *et al.*, Prebiotic effects and intestinal fermentation of cereal arabinoxylans and arabinoxylan oligosaccharides in rats depend strongly on their structural properties and joint presence. *Mol. Nutr. Food Res.* **55**, 1862-1874 (2011).
273. M. Mendis, S. Simsek, Arabinoxylans and human health. *Food Hydrocolloid* **42**, 239-243 (2014).
274. J. M. Lattimer, M. D. Haub, Effects of dietary fiber and its components on metabolic health. *Nutrients* **2**, 1266-1289 (2010).
275. A. G. Adam-Perrot, L.; Sanders, L.; Bouvier, S.; Combe, C.; Van Den Abbeele, R.; Potter, S.; Einerhand, A. W. C., in *Prebiotics and Probiotics Science and Technology, Volume 1,* D. Charalampopoulos, R. A. Rastall, Eds. (Springer-Verlag, New York, 2009).
276. S. Tachon, J. N. Zhou, M. Keenan, R. Martin, M. L. Marco, The intestinal microbiota in aged mice is modulated by dietary resistant starch and correlated with improvements in host responses. *FEMS Microbiol. Ecol.* **83**, 299-309 (2013).
277. I. Martinez, J. Kim, P. R. Duffy, V. L. Schlegel, J. Walter, Resistant Starches Types 2 and 4 Have Differential Effects on the Composition of the Fecal Microbiota in Human Subjects. *Plos One* **5**, e15046 (2010).
278. A. W. Walker *et al.*, Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *Isme J* **5**, 220-230 (2011).
279. S. Duranti *et al.*, Genomic Characterization and Transcriptional Studies of the Starch-Utilizing Strain Bifidobacterium adolescentis 22L. *Appl. Environ. Microbiol.* **80**, 6080-6090 (2014).
280. M. O. Motherway *et al.*, Characterization of ApuB, an extracellular type II amylopullulanase from Bifidobacterium breve UCC2003. *Appl. Environ. Microbiol.* **74**, 6271-6279 (2008).
281. D. H. Jung *et al.*, Complete genome sequence of Bifidobacterium choerinum FMB-1, a resistant starch-degrading bacterium. *J. Biotechnol.* **274**, 28-32 (2018).
282. K. M. Behall, D. J. Scholfield, J. G. Hallfrisch, H. G. M. Liljeberg-Elmstahl, Consumption of both resistant starch and beta-glucan improves postprandial plasma glucose and insulin in women. *Diabetes Care* **29**, 976-981 (2006).
283. A. C. Nilsson, E. M. Ostman, J. J. Hoist, I. M. E. Bjorck, Including indigestible carbohydrates in the evening meal of healthy subjects improves glucose tolerance, lowers inflammatory markers, and increases satiety after a subsequent standardized breakfast. *J. Nutr.* **138**, 732-739 (2008).

284. M. P. Maziarz *et al.*, Resistant starch lowers postprandial glucose and leptin in overweight adults consuming a moderate-to-high-fat diet: a randomized-controlled trial. *Nutr J* **16**, 14 (2017).
285. P. Louis, K. P. Scott, S. H. Duncan, H. J. Flint, Understanding the effects of diet on bacterial metabolism in the large intestine. *J. Appl. Microbiol.* **102**, 1197-1208 (2007).
286. G. A. Lugli *et al.*, Tracking the taxonomy of the genus Bifidobacterium based on a phylogenomic approach. *Appl. Environ. Microbiol.* **84**, e02249-02217 (2017).
287. G. A. Lugli *et al.*, Comparative genomic and phylogenomic analyses of the Bifidobacteriaceae family. *BMC Genomics* **18**, (2017).
288. T. Vatanen *et al.*, Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. *Nat Microbiol* **4**, 470-479 (2019).
289. T. Odamaki *et al.*, Genomic diversity and distribution of Bifidobacterium longum subsp longum across the human lifespan. *Sci Rep-Uk* **8**, 85 (2018).
290. K. James, M. O. Motherway, C. Penno, R. L. O'Brien, D. van Sinderen, Bifidobacterium breve UCC2003 employs multiple transcriptional regulators to control metabolism of particular human milk oligosaccharides. *Appl. Environ. Microbiol.* **84**, e02774-02717 (2018).
291. E. Ozcan, D. A. Sela, Inefficient Metabolism of the Human Milk Oligosaccharides Lacto-N-tetraose and Lacto-N-neotetraose Shifts Bifidobacterium longum subsp. infantis Physiology. *Front Nutr* **5**, 46 (2018).
292. G. A. Lugli *et al.*, Dissecting the Evolutionary Development of the Species Bifidobacterium animalis through Comparative Genomics Analyses. *Appl. Environ. Microbiol.* **85**, e02806-02818 (2019).
293. G. A. Lugli *et al.*, Unveiling Genomic Diversity among Members of the Species Bifidobacterium pseudolongum, a Widely Distributed Gut Commensal of the Animal Kingdom. *Appl. Environ. Microbiol.* **85**, e03065-03018 (2019).
294. C. Milani *et al.*, Unveiling bifidobacterial biogeography across the mammalian branch of the tree of life. *Isme J* **11**, 2834-2847 (2017).
295. G. A. Lugli *et al.*, Evolutionary development and co-phylogeny of primate-associated bifidobacteria. *Environ. Microbiol.* **22**, 3375-3393 (2020).
296. L. C. Roger, A. L. McCartney, Longitudinal investigation of the faecal microbiota of healthy full-term infants using fluorescence in situ hybridization and denaturing gradient gel electrophoresis. *Microbiol-Sgm* **156**, 3317-3328 (2010).
297. A. Stradiotto *et al.*, Spatial Organization of the Yellow-Necked Mouse: Effects of Density and Resource Availability. *J. Mammal.* **90**, 704-714 (2009).
298. S. Even, C. Garrigues, P. Loubiere, N. D. Lindley, M. Cocaign-Bousquet, Pyruvate Metabolism in Lactococcus lactis Is Dependent upon Glyceraldehyde-3-phosphate Dehydrogenase Activity. *Metab. Eng.* **1**, 198-205 (1999).
299. C. Solem, B. J. Koebmann, P. R. Jensen, Glyceraldehyde-3-phosphate dehydrogenase has no control over glycolytic flux in Lactococcus lactis MG1363. *J. Bacteriol.* **185**, 1564-1571 (2003).
300. W. G. Weisburg, S. M. Barns, D. A. Pelletier, D. J. Lane, 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.* **173**, 697-703 (1991).
301. A. B. Yoo, M. A. Jette, M. Grondona, SLURM: Simple linux utility for resource management. *Job Scheduling Strategies for Parallel Processing* **2862**, 44-60 (2003).
302. D. R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821-829 (2008).
303. A. Bankevich *et al.*, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455-477 (2012).

304. D. E. Wood, S. L. Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15**, R46 (2014).
305. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890 (2018).
306. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).
307. L. Pritchard, R. H. Glover, S. Humphris, J. G. Elphinstone, I. K. Toth, Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods-Uk* **8**, 12-24 (2016).
308. J. Chun *et al.*, Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int. J. Syst. Evol. Microbiol.* **68**, 461-466 (2018).
309. M. R. Olm *et al.*, Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different in situ growth rates. *Genome Res.* **27**, 601-612 (2017).
310. T. Seemann, Barrnap: BAsic Rapid Ribosomal RNA Predictor. Available at: https://github.com/tseemann/barrnap (2014).
311. E. Pruesse, J. Peplies, F. O. Glockner, SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823-1829 (2012).
312. S. Capella-Gutierrez, J. M. Silla-Martinez, T. Gabaldon, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
313. D. Darriba, G. L. Taboada, R. Doallo, D. Posada, jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**, 772 (2012).
314. C. M. Hurvich, C. Tsai, A corrected Akaike Information Criterion for vector autoregressive model selection. *Journal of Time Series Analysis* **14**, 271-279 (1993).
315. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2-Approximately Maximum-Likelihood Trees for Large Alignments. *Plos One* **5**, e9490 (2010).
316. K. Katoh, D. M. Standley, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).
317. J. Qi, H. Luo, B. Hao, CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Res.* **32**, W45-47 (2004).
318. A. J. Page *et al.*, Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-3693 (2015).
319. K. Katoh, J. Rozewicki, K. D. Yamada, MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform* **20**, 1160-1166 (2019).
320. T. Seemann, snippy: fast bacterial variant calling from NGS reads. Available at: https://github.com/tseemann/snippy (2015).
321. N. J. Croucher *et al.*, Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).
322. G. Talavera, J. Castresana, Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564-577 (2007).
323. A. J. Page *et al.*, SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genomics* **2**, e000056 (2016).
324. T. Seemann, A. J. Page, F. Klotzl, snp-dists: Pairwise SNP distance matrix from a FASTA sequence alignment. Available at: https://github.com/tseemann/snp-dists (2017).
325. O. Brynildsrud, J. Bohlin, L. Scheffer, V. Eldholm, Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol* **17**, 238 (2016).

326. A. M. Eren *et al.*, Anvi'o: an advanced analysis and visualization platform for 'omics data. *Peerj* **3**, e1319 (2015).
327. L. T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268-274 (2015).
328. S. Whelan, N. Goldman, A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691-699 (2001).
329. P. Legendre, Y. Desdevises, E. Bazin, A statistical test for host-parasite coevolution. *Syst. Biol.* **51**, 217-234 (2002).
330. E. Paradis, K. Schliep, ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526-528 (2019).
331. J. Huerta-Cepas *et al.*, Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115-2122 (2017).
332. H. Zhang *et al.*, dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **46**, W95-W101 (2018).
333. M. Csuros, I. Miklos, A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. *Lect Notes Comput Sc* **3909**, 206-220 (2006).
334. C. Camacho *et al.*, BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
335. S. Waack *et al.*, Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics* **7**, 142 (2006).
336. C. Bertelli *et al.*, IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res.* **45**, W30-W35 (2017).
337. D. Couvin *et al.*, CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* **46**, W246-W251 (2018).
338. A. Biswas, R. H. Staals, S. E. Morales, P. C. Fineran, C. M. Brown, CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* **17**, 356 (2016).
339. M. B. Dion, S. J. Labrie, S. A. Shah, S. Moineau, CRISPRStudio: A User-Friendly Software for Rapid CRISPR Array Visualization. *Viruses* **10**, 602 (2018).
340. S. Roux, F. Enault, B. L. Hurwitz, M. B. Sullivan, VirSorter: mining viral signal from microbial genomic data. *Peerj* **3**, e985 (2015).
341. N. F. Alikhan, N. K. Petty, N. L. Ben Zakour, S. A. Beatson, BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **12**, 402 (2011).
342. M. O. Arntzen, I. L. Karlskas, M. Skaugen, V. G. H. Eijsink, G. Mathiesen, Proteomic Investigation of the Response of Enterococcus faecalis V583 when Cultivated in Urine. *Plos One* **10**, e0126694 (2015).
343. J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367-1372 (2008).
344. J. Cox *et al.*, Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Molecular & Cellular Proteomics* **13**, 2513-2526 (2014).
345. S. Tyanova *et al.*, The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731-740 (2016).
346. I. Letunic, P. Bork, Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242-245 (2016).

347. R. C. Team, R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Available at: http://www.r-project.org/index.html (2017).
348. L. Wampach *et al.*, Colonization and Succession within the Human Gut Microbiome by Archaea, Bacteria, and Microeukaryotes during the First Year of Life. *Front Microbiol* **8**, 738 (2017).
349. M. G. de Aguero *et al.*, The maternal microbiota drives early postnatal innate immune development. *Science* **351**, 1296-1301 (2016).
350. A. Sivan *et al.*, Commensal Bifidobacterium promotes antitumor immunity and facilitates anti-PD-L1 efficacy. *Science* **350**, 1084-1089 (2015).
351. M. Aaboud *et al.*, Search for High-Mass Resonances Decaying to taunu in pp Collisions at sqrt[s]=13 TeV with the ATLAS Detector. *Phys Rev Lett* **120**, 161802 (2018).
352. T. Thongaram, J. L. Hoeflinger, J. Chow, M. J. Miller, Human milk oligosaccharide consumption by probiotic and human-associated bifidobacteria and lactobacilli. *J. Dairy Sci.* **100**, 7825-7833 (2017).
353. S. Dogra *et al.*, Dynamics of infant gut microbiota are influenced by delivery mode and gestational duration and are associated with subsequent adiposity. *mBio* **6**, e02419-02414 (2015).
354. Y. Shao *et al.*, Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117-121 (2019).
355. J. D. Forbes *et al.*, Association of Exposure to Formula in the Hospital and Subsequent Infant Feeding Practices With Gut Microbiota and Risk of Overweight in the First Year of Life. *JAMA pediatrics* **172**, e181161 (2018).
356. M. X. Maldonado-Gomez *et al.*, Stable Engraftment of Bifidobacterium longum AH1206 in the Human Gut Depends on Individualized Features of the Resident Microbiome. *Cell Host Microbe* **20**, 515-526 (2016).
357. K. Oki *et al.*, Long-term colonization exceeding six years from early infancy of Bifidobacterium longum subsp. longum in human gut. *BMC Microbiol.* **18**, 209 (2018).
358. P. Mattarelli, C. Bonaparte, B. Pot, B. Biavati, Proposal to reclassify the three biotypes of Bifidobacterium longum as three subspecies: Bifidobacterium longum subsp. longum subsp. nov., Bifidobacterium longum subsp. infantis comb. nov. and Bifidobacterium longum subsp. suis comb. nov. *Int. J. Syst. Evol. Microbiol.* **58**, 767-772 (2008).
359. E. Yanokura *et al.*, Subspeciation of Bifidobacterium longum by multilocus approaches and amplified fragment length polymorphism: Description of B. longum subsp. suillum subsp. nov., isolated from the faeces of piglets. *Syst. Appl. Microbiol.* **38**, 305-314 (2015).
360. D. Garrido, D. Barile, D. A. Mills, A molecular basis for bifidobacterial enrichment in the infant gastrointestinal tract. *Adv Nutr* **3**, 415S-421S (2012).
361. F. Magne *et al.*, A longitudinal study of infant faecal microbiota during weaning. *FEMS Microbiol. Ecol.* **58**, 563-571 (2006).
362. C. Palmer, E. M. Bik, D. B. DiGiulio, D. A. Relman, P. O. Brown, Development of the human infant intestinal microbiota. *PLoS Biol.* **5**, e177 (2007).
363. L. C. Roger, A. Costabile, D. T. Holland, L. Hoyles, A. L. McCartney, Examination of faecal Bifidobacterium populations in breast- and formula-fed infants during the first 18 months of life. *Microbiol-Sgm* **156**, 3329-3341 (2010).
364. A. V. Chaplin *et al.*, Intraspecies Genomic Diversity and Long-Term Persistence of Bifidobacterium longum. *Plos One* **10**, e0135658 (2015).

365. S. Arboleya *et al.*, Gene-trait matching across the Bifidobacterium longum pan-genome reveals considerable diversity in carbohydrate catabolism among human infant strains. *BMC Genomics* **19**, 33 (2018).
366. H. Ichinose, M. Yoshida, Z. Fujimoto, S. Kaneko, Characterization of a modular enzyme of exo-1,5-alpha-L-arabinofuranosidase and arabinan binding module from Streptomyces avermitilis NBRC14893. *Appl. Microbiol. Biotechnol.* **80**, 399-408 (2008).
367. S. Ahmed *et al.*, A novel alpha-L-arabinofuranosidase of family 43 glycoside hydrolase (Ct43Araf) from Clostridium thermocellum. *Plos One* **8**, e73575 (2013).
368. C. Milani *et al.*, Exploring Vertical Transmission of Bifidobacteria from Mother to Child. *Appl Environ Microbiol* **81**, 7078-7087 (2015).
369. D. A. Sela, D. A. Mills, Nursing our microbiota: molecular linkages between bifidobacteria and milk oligosaccharides. *Trends Microbiol* **18**, 298-307 (2010).
370. A. H. Viborg *et al.*, Biochemical and kinetic characterisation of a novel xylooligosaccharide-upregulated GH43 beta-D-xylosidase/alpha-L-arabinofuranosidase (BXA43) from the probiotic Bifidobacterium animalis subsp lactis BB-12. *Amb Express* **3**, 56 (2013).
371. H. R. Sorensen, S. Pedersen, C. T. Jorgensen, A. S. Meyer, Enzymatic hydrolysis of wheat arabinoxylan by a recombinant "minimal" enzyme cocktail containing beta-xylosidase and novel endo-1,4-beta-xylanase and alpha-l-arabinofuranosidase activities. *Biotechnol. Prog.* **23**, 100-107 (2007).
372. S. K. Kang *et al.*, Three forms of thermostable lactose-hydrolase from Thermus sp IB-21: cloning, expression, and enzyme characterization. *J. Biotechnol.* **116**, 337-346 (2005).
373. S. W. A. Hinz, L. A. M. van den Broek, G. Beldman, J. P. Vincken, A. G. J. Voragen, Beta-galactosidase from Bifidobacterium adolescentis DSM20083 prefers beta(1,4)-galactosides over lactose. *Appl. Microbiol. Biotechnol.* **66**, 276-284 (2004).
374. H. Taniguchi, Y. Honnda, in *Encyclopedia of Microbiology,* M. Schaechter, Ed. (Elsevier Inc., San Diego, 2009), pp. 159-173.
375. J. Intra, G. Pavesi, D. S. Horner, Phylogenetic analyses suggest multiple changes of substrate specificity within the glycosyl hydrolase 20 family. *BMC Evol. Biol.* **8**, 214 (2008).
376. H. Suzuki, A. Murakami, K. Yoshida, Motif-guided identification of a glycoside hydrolase family 1 alpha-L-arabinofuranosidase in Bifidobacterium adolescentis. *Biosci Biotechnol Biochem* **77**, 1709-1714 (2013).
377. S. Hwang, K. H. Choi, J. Kim, J. Cha, Biochemical characterization of 4-alpha-glucanotransferase from Saccharophagus degradans 2-40 and its potential role in glycogen degradation. *FEMS Microbiol. Lett.* **344**, 145-151 (2013).
378. D. Garrido *et al.*, Comparative transcriptomics reveals key differences in the response to milk oligosaccharides of infant gut-associated bifidobacteria. *Sci Rep* **5**, 13517 (2015).
379. D. A. Sela *et al.*, Bifidobacterium longum subsp. infantis ATCC 15697 alpha-fucosidases are active on fucosylated human milk oligosaccharides. *Appl. Environ. Microbiol.* **78**, 795-803 (2012).
380. M. Kitaoka, Bifidobacterial enzymes involved in the metabolism of human milk oligosaccharides. *Adv Nutr* **3**, 422S-429S (2012).
381. S. W. Hinz, M. I. Pastink, L. A. van den Broek, J. P. Vincken, A. G. Voragen, Bifidobacterium longum endogalactanase liberates galactotriose from type I galactans. *Appl. Environ. Microbiol.* **71**, 5501-5510 (2005).
382. A. S. Luis *et al.*, Dietary pectic glycans are degraded by coordinated enzyme pathways in human colonic Bacteroides. *Nat Microbiol* **3**, 210-219 (2018).

383. L. A. Miosge *et al.*, Comparison of predicted and actual consequences of missense mutations. *Proc Natl Acad Sci U S A* **112**, E5189-5198 (2015).
384. S. Mills, C. Stanton, J. A. Lane, G. J. Smith, R. P. Ross, Precision Nutrition and the Microbiome, Part I: Current State of the Science. *Nutrients* **11**, 923 (2019).
385. J. M. Andersen *et al.*, Transcriptional analysis of oligosaccharide utilization by Bifidobacterium lactis Bl-04. *BMC Genomics* **14**, 312 (2013).
386. S. Yan *et al.*, Functional and structural characterization of a beta-glucosidase involved in saponin metabolism from intestinal bacteria. *Biochem. Biophys. Res. Commun.* **496**, 1349-1356 (2018).
387. T. Pozzo, J. L. Pasten, E. N. Karlsson, D. T. Logan, Structural and functional analyses of beta-glucosidase 3B from Thermotoga neapolitana: a thermostable three-domain representative of glycoside hydrolase 3. *J. Mol. Biol.* **397**, 724-739 (2010).
388. R. N. Florindo *et al.*, Structural and biochemical characterization of a GH3 beta-glucosidase from the probiotic bacteria Bifidobacterium adolescentis. *Biochimie* **148**, 107-115 (2018).
389. F. Asnicar *et al.*, Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems* **2**, e00164-00116 (2017).
390. K. Mikami, M. Kimura, H. Takahashi, Influence of maternal bifidobacteria on the development of gut bifidobacteria in infants. *Pharmaceuticals (Basel)* **5**, 629-642 (2012).
391. S. Lax *et al.*, Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science* **345**, 1048-1052 (2014).
392. J. Dworkin, R. Losick, Linking nutritional status to gene activation and development. *Genes Dev.* **15**, 1051-1054 (2001).
393. J. Slager, J. W. Veening, Hard-Wired Control of Bacterial Processes by Chromosomal Gene Location. *Trends Microbiol.* **24**, 788-800 (2016).
394. D. Rios-Covian *et al.*, Interactions between Bifidobacterium and Bacteroides species in cofermentations are affected by carbon sources, including exopolysaccharides produced by bifidobacteria. *Appl. Environ. Microbiol.* **79**, 7518-7524 (2013).
395. C. Schwab *et al.*, Trophic Interactions of Infant Bifidobacteria and Eubacterium hallii during L-Fucose and Fucosyllactose Degradation. *Front Microbiol* **8**, 95 (2017).
396. S. Parche *et al.*, Sugar transport systems of Bifidobacterium longum NCC2705. *J. Mol. Microbiol. Biotechnol.* **12**, 9-19 (2007).
397. D. Liu *et al.*, Proteomics analysis of Bifidobacterium longum NCC2705 growing on glucose, fructose, mannose, xylose, ribose, and galactose. *Proteomics* **11**, 2628-2638 (2011).
398. C. Ferrario *et al.*, Modulation of the eps-ome transcription of bifidobacteria through simulation of human intestinal environment. *FEMS Microbiol. Ecol.* **92**, fiw056 (2016).
399. B. D. Muegge *et al.*, Diet Drives Convergence in Gut Microbiome Functions Across Mammalian Phylogeny and Within Humans. *Science* **332**, 970-974 (2011).
400. T. A. Suzuki, Links between Natural Variation in the Microbiome and Host Fitness in Wild Mammals. *Integr. Comp. Biol.* **57**, 756-769 (2017).
401. M. McFall-Ngai *et al.*, Animals in a bacterial world, a new imperative for the life sciences. *Proc Natl Acad Sci U S A* **110**, 3229-3236 (2013).
402. B. K. Trevelline, S. S. Fontaine, B. K. Hartup, K. D. Kohl, Conservation biology needs a microbial renaissance: a call for the consideration of host-associated microbiota in wildlife management practices. *Proc Biol Sci* **286**, 20182448 (2019).
403. S. Bahrndorff, T. Alemu, T. Alemneh, J. Lund Nielsen, The Microbiome of Animals: Implications for Conservation Biology. *Int J Genomics* **2016**, 5304028 (2016).

404. A. H. Moeller *et al.*, Cospeciation of gut microbiota with hominids. *Science* **353**, 380-382 (2016).
405. M. Groussin *et al.*, Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. *Nat Commun* **8**, 14319 (2017).
406. C. A. Gaulke *et al.*, Ecophylogenetics Clarifies the Evolutionary Association between Mammals and Their Gut Microbiota. *mBio* **9**, e01348-01318 (2018).
407. N. D. Youngblut *et al.*, Host diet and evolutionary history explain different aspects of gut microbiome diversity among vertebrate clades. *Nat Commun* **10**, 2200 (2019).
408. C. A. Lozupone *et al.*, The convergence of carbohydrate active gene repertoires in human gut microbes. *Proc Natl Acad Sci U S A* **105**, 15076-15081 (2008).
409. K. Makarova *et al.*, Comparative genomics of the lactic acid bacteria. *Proc Natl Acad Sci U S A* **103**, 15611-15616 (2006).
410. K. R. Foster, J. Schluter, K. Z. Coyte, S. Rakoff-Nahoum, The evolution of the host microbiome as an ecosystem on a leash. *Nature* **548**, 43-51 (2017).
411. S. van Vliet, M. Doebeli, The role of multilevel selection in host microbiome evolution. *Proc Natl Acad Sci U S A* **116**, 20591-20597 (2019).
412. M. Groussin, F. Mazel, E. J. Alm, Co-evolution and Co-speciation of Host-Gut Bacteria Systems. *Cell Host Microbe* **28**, 12-22 (2020).
413. S. Duranti *et al.*, Characterization of the phylogenetic diversity of five novel species belonging to the genus Bifidobacterium: Bifidobacterium castoris sp. nov., Bifidobacterium callimiconis sp. nov., Bifidobacterium goeldii sp. nov., Bifidobacterium samirii sp. nov. and Bifidobacterium dolichotidis sp. nov. *Int. J. Syst. Evol. Microbiol.* **69**, 1288-1298 (2019).
414. P. Schumann, Peptidoglycan Structure. *Method Microbiol* **38**, 101-129 (2011).
415. B. Snel, P. Bork, M. A. Huynen, Genome phylogeny based on gene content. *Nat. Genet.* **21**, 108-110 (1999).
416. B. E. Dutilh, M. A. Huynen, W. J. Bruno, B. Snel, The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *J. Mol. Evol.* **58**, 527-539 (2004).
417. J. R. Michaux, P. Chevret, M. G. Filippucci, M. Macholan, Phylogeny of the genus Apodemus with a special emphasis on the subgenus Sylvaemus using the nuclear IRBP gene and two mitochondrial markers: cytochrome b and 12S rRNA. *Mol Phylogenet Evol* **23**, 123-136 (2002).
418. B. Henrissat, G. J. Davies, Glycoside hydrolases and glycosyltransferases. Families, modules, and implications for genomics. *Plant Physiol.* **124**, 1515-1519 (2000).
419. M. R. Stam, E. G. Danchin, C. Rancurel, P. M. Coutinho, B. Henrissat, Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins. *Protein Eng. Des. Sel.* **19**, 555-562 (2006).
420. T. Miyazaki, Y. Ishizaki, M. Ichikawa, A. Nishikawa, T. Tonozuka, Structural and biochemical characterization of novel bacterial alpha-galactosidases belonging to glycoside hydrolase family 31. *Biochem. J.* **469**, 145-158 (2015).
421. M. A. Hachem *et al.*, Raffinose family oligosaccharide utilisation by probiotic bacteria: insight into substrate recognition, molecular architecture and diversity of GH36 α-galactosidases. *Biocatalysis Biotransformation* **30**, 316-325 (2012).
422. P. Viens, M. E. Lacombe-Harvey, R. Brzezinski, Chitosanases from Family 46 of Glycoside Hydrolases: From Proteins to Phenotypes. *Mar. Drugs* **13**, 6566-6587 (2015).
423. K. Fujita, Y. Takashi, E. Obuchi, K. Kitahara, T. Suganuma, Characterization of a novel beta-L-arabinofuranosidase in Bifidobacterium longum: functional

elucidation of a DUF1680 protein family member. *J. Biol. Chem.* **289**, 5240-5249 (2014).
424. C. Breton, L. Snajdrova, C. Jeanneau, J. Koca, A. Imberty, Structures and mechanisms of glycosyltransferases. *Glycobiology* **16**, 29R-37R (2006).
425. C. Hidalgo-Cantabrana *et al.*, Genomic Overview and Biological Functions of Exopolysaccharide Biosynthesis in Bifidobacterium spp. *Appl. Environ. Microbiol.* **80**, 9-18 (2014).
426. B. L. Cantarel *et al.*, The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* **37**, D233-238 (2009).
427. H. Liu, W. Ren, M. Ly, H. Li, S. Wang, Characterization of an Alkaline GH49 Dextranase from Marine Bacterium Arthrobacter oxydans KQ11 and Its Application in the Preparation of Isomalto-Oligosaccharide. *Mar. Drugs* **17**, 479 (2019).
428. H. Michlmayr *et al.*, Arabinoxylan oligosaccharide hydrolysis by family 43 and 51 glycosidases from Lactobacillus brevis DSM 20054. *Appl. Environ. Microbiol.* **79**, 6747-6754 (2013).
429. C. Roca, V. D. Alves, F. Freitas, M. A. Reis, Exopolysaccharides enriched in rare sugars: bacterial sources, production, and applications. *Front Microbiol* **6**, 288 (2015).
430. S. Balzaretti *et al.*, A Novel Rhamnose-Rich Hetero-exopolysaccharide Isolated from Lactobacillus paracasei DG Activates THP-1 Human Monocytic Cells. *Appl. Environ. Microbiol.* **83**, e02702-02716 (2017).
431. F. Hille *et al.*, The Biology of CRISPR-Cas: Backward and Forward. *Cell* **172**, 1239-1259 (2018).
432. J. K. Nunez *et al.*, Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528-534 (2014).
433. I. Grissa, G. Vergnaud, C. Pourcel, CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* **35**, W52-57 (2007).
434. A. E. Briner *et al.*, Occurrence and Diversity of CRISPR-Cas Systems in the Genus Bifidobacterium. *Plos One* **10**, e0133661 (2015).
435. C. Hidalgo-Cantabrana, A. B. Crawley, B. Sanchez, R. Barrangou, Characterization and Exploitation of CRISPR Loci in Bifidobacterium longum. *Front Microbiol* **8**, 1851 (2017).
436. G. Wang *et al.*, The Diversity of the CRISPR-Cas System and Prophages Present in the Genome Reveals the Co-evolution of Bifidobacterium pseudocatenulatum and Phages. *Front Microbiol* **11**, 1088 (2020).
437. G. A. Lugli *et al.*, Prophages of the genus Bifidobacterium as modulating agents of the infant gut microbiota. *Environ. Microbiol.* **18**, 2196-2213 (2016).
438. D. Botstein, A theory of modular evolution for bacteriophages. *Ann. N. Y. Acad. Sci.* **354**, 484-490 (1980).
439. J. R. Michaux, R. Libois, M.-G. Filipucci, So close and so different: comparative phylogeography of two small mammal species, the Yellow-necked fieldmouse (Apodemus flavicollis) and the Woodmouse (Apodemus sylvaticus) in the Western Palearctic region. *Heredity* **94**, 52-63 (2005).
440. D. Ge *et al.*, Evolutionary history of field mice (Murinae: Apodemus), with emphasis on morphological variation among species in China and description of a new species. *Zool. J. Linn. Soc.* **187**, 5188-5534 (2019).
441. A. H. Moeller *et al.*, Sympatric chimpanzees and gorillas harbor convergent gut microbial communities. *Genome Res.* **23**, 1715-1720 (2013).

442. S. R. Bogatyrev, J. C. Rolando, R. F. Ismagilov, Self-reinoculation with fecal flora changes microbiota density and composition leading to an altered bile-acid profile in the mouse small intestine. *Microbiome* **8**, 19 (2020).
443. S. C. L. Knowles, R. M. Eccles, L. Baltrunaite, Species identity dominates over environment in shaping the microbiota of small mammals. *Ecol. Lett.* **22**, 826-837 (2019).
444. C. H. S. Watts, The foods eaten by wood mice (Apodemus sylvaticus) and bank voles (Clethrionomys glareolus) in Wytham Woods, Berkshire. *J. Anim. Ecol.* **37**, 25-41 (1968).
445. K. F. Abt, Bock, Seasonal variations of diet composition in farmland field mice Apodemus spp. and bank voles Clethrionomys glareolus. *Acta Theriologica* **43**, 379-389 (1998).
446. L. M. Rogers, M. L. Gorman, The diet of the wood mouse Apodemus sylvaticus on set-aside land. *J. Zool.* **235**, 77-83 (1995).
447. K. M. Van Laere, G. Beldman, A. G. Voragen, A new arabinofuranohydrolase from Bifidobacterium adolescentis able to remove arabinosyl residues from double-substituted xylose units in arabinoxylan. *Appl. Microbiol. Biotechnol.* **47**, 231-235 (1997).
448. A. Margolles, C. G. de los Reyes-Gavilan, Purification and functional characterization of a novel alpha-L-arabinofuranosidase from Bifidobacterium longum B667. *Appl. Environ. Microbiol.* **69**, 5096-5103 (2003).
449. M. Kataržytė, E. Kutorga, Small mammal mycophagy in hemiboreal forest communities of Lithuania. *Central European Journal of Biology* **6**, 446-456 (2011).
450. H. W. Lee, Y. S. Park, J. S. Jung, W. S. Shin, Chitosan oligosaccharides, dp 2-8, have prebiotic effect on the Bifidobacterium bifidium and Lactobacillus sp. *Anaerobe* **8**, 319-324 (2002).
451. C. L. Vernazza, G. R. Gibson, R. A. Rastall, In vitro fermentation of chitosan derivatives by mixed cultures of human faecal bacteria. *Carbohyd Polym*, 539–545 (2005).
452. C. M. Yang *et al.*, Effect of chito-oligosaccharide on growth performance, intestinal barrier function, intestinal morphology and cecal microflora in weaned pigs. *J. Anim. Sci.* **90**, 2671-2676 (2012).
453. C. Zhang, S. Jiao, Z. A. Wang, Y. Du, Exploring Effects of Chitosan Oligosaccharides on Mice Gut Microbiota in in vitro Fermentation and Animal Model. *Front Microbiol* **9**, 2388 (2018).
454. P. A. Prieto *et al.*, Remodeling of mouse milk glycoconjugates by transgenic expression of a human glycosyltransferase. *J. Biol. Chem.* **270**, 29515-29519 (1995).
455. J. Audy, S. Labrie, D. Roy, G. Lapointe, Sugar source modulates exopolysaccharide biosynthesis in Bifidobacterium longum subsp. longum CRC 002. *Microbiology* **156**, 653-664 (2010).
456. N. Salazar, M. Gueimonde, A. M. Hernandez-Barranco, P. Ruas-Madiedo, C. G. de los Reyes-Gavilan, Exopolysaccharides produced by intestinal Bifidobacterium strains act as fermentable substrates for human intestinal bacteria. *Appl. Environ. Microbiol.* **74**, 4737-4745 (2008).
457. D. Pungel *et al.*, Bifidobacterium breve UCC2003 Exopolysaccharide Modulates the Early Life Microbiota by Acting as a Potential Dietary Substrate. *Nutrients* **12**, 948 (2020).
458. K. S. Makarova *et al.*, An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13**, 722-736 (2015).
459. P. Hyman, S. T. Abedon, Bacteriophage host range and bacterial resistance. *Adv. Appl. Microbiol.* **70**, 217-248 (2010).

460. C. Canchaya, C. Proux, G. Fournous, A. Bruttin, H. Brussow, Prophage genomics. *Microbiol. Mol. Biol. Rev.* **67**, 238-276 (2003).
461. M. Asadulghani *et al.*, The defective prophage pool of Escherichia coli O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants. *PLoS Pathog* **5**, e1000408 (2009).
462. L. M. Bobay, M. Touchon, E. P. Rocha, Pervasive domestication of defective prophages by bacteria. *Proc Natl Acad Sci U S A* **111**, 12127-12132 (2014).
463. F. Dini-Andreote, F. D. Andreote, W. L. Araujo, J. T. Trevors, J. D. van Elsas, Bacterial genomes: habitat specificity and uncharted organisms. *Microb. Ecol.* **64**, 1-7 (2012).
464. M. Modesto *et al.*, Bifidobacterium callitrichidarum sp. nov. from the faeces of the emperor tamarin (Saguinus imperator). *Int. J. Syst. Evol. Microbiol.* **68**, 141-148 (2018).
465. M. Modesto *et al.*, Bifidobacterium catulorum sp. nov., a novel taxon from the faeces of the baby common marmoset (Callithrix jacchus). *Int. J. Syst. Evol. Microbiol.* **68**, 575-581 (2018).
466. A. O'Callaghan, F. Bottacini, M. O. Motherway, D. van Sinderen, Pangenome analysis of Bifidobacterium longum and site-directed mutagenesis through by-pass of restriction-modification systems. *BMC Genomics* **16**, 832 (2015).
467. Y. Zhang, S. M. Sievert, Pan-genome analyses identify lineage- and niche-specific markers of evolution and adaptation in Epsilonproteobacteria. *Front Microbiol* **5**, 110 (2014).
468. R. G. Leuschner, J. Bew, P. Simpson, P. R. Ross, C. Stanton, A collaborative study of a method for the enumeration of probiotic bifidobacteria in animal feed. *Int. J. Food Microbiol.* **83**, 161-170 (2003).
469. F. J. Munoa, R. Pares, Selective medium for isolation and enumeration of Bifidobacterium spp. *Appl. Environ. Microbiol.* **54**, 1715-1718 (1988).
470. L. Arroyo, Cotton, L.N., Martin, J.H., AMC agar - a composite medium for selective enumeration of Bifidobacterium longum *Cultured Dairy Products Journal* **30**, 12-15 (1995).
471. S. Silvi, C. J. Rumney, I. R. Rowland, An assessment of three selective media for bifidobacteria in faeces. *J Appl Bacteriol* **81**, 561-564 (1996).
472. S. Ingham, Use of modified Lactobacillus selective medium and Bifidobacterium iodoacetate medium for differential enumeration of Lactobacillus acidophilus and Bifidobacterium spp. in powered nutritional products. *J. Food Prot.* **62**, 77-80 (1999).
473. C. Garcia-Aljaro *et al.*, Neoscardovia arbecensis gen. nov., sp. nov., isolated from porcine slurries. *Syst. Appl. Microbiol.* **35**, 374-379 (2012).
474. M. Cobo-Simon, J. Tamames, Relating genomic characteristics to environmental preferences and ubiquity in different microbial taxa. *BMC Genomics* **18**, 499 (2017).
475. E. Litchman, Invisible invaders: non-pathogenic invasive microbes in aquatic and terrestrial ecosystems. *Ecol. Lett.* **13**, 1560-1572 (2010).
476. C. L. Dupont *et al.*, Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *Isme J* **6**, 1186-1199 (2012).
477. J. Raes, J. O. Korbel, M. J. Lercher, C. von Mering, P. Bork, Prediction of effective genome size in metagenomic samples. *Genome Biol* **8**, R10 (2007).
478. F. E. Angly *et al.*, The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol* **5**, e1000593 (2009).
479. J. A. Ranea, D. W. Buchan, J. M. Thornton, C. A. Orengo, Evolution of protein superfamilies and bacterial genome size. *J. Mol. Biol.* **336**, 871-887 (2004).

480. J. Tamames, P. D. Sanchez, P. I. Nikel, C. Pedros-Alio, Quantifying the Relative Importance of Phylogeny and Environmental Preferences As Drivers of Gene Content in Prokaryotic Microorganisms. *Front Microbiol* **7**, 433 (2016).
481. J. P. McCutcheon, N. A. Moran, Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10**, 13-26 (2011).
482. Y. I. Wolf, E. V. Koonin, Genome reduction as the dominant mode of evolution. *Bioessays* **35**, 829-837 (2013).
483. H. Toh *et al.*, Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of Sodalis glossinidius in the tsetse host. *Genome Res.* **16**, 149-156 (2006).
484. G. R. Burke, N. A. Moran, Massive genomic decay in Serratia symbiotica, a recently evolved symbiont of aphids. *Genome Biol Evol* **3**, 195-208 (2011).
485. N. Tao *et al.*, Evolutionary glycomics: characterization of milk oligosaccharides in primates. *J Proteome Res* **10**, 1548-1557 (2011).
486. F. J. Lee, D. B. Rusch, F. J. Stewart, H. R. Mattila, I. L. Newton, Saccharide breakdown and fermentation by the honey bee gut microbiome. *Environ. Microbiol.* **17**, 796-815 (2015).
487. R. Lamendella, J. W. Santo Domingo, C. Kelty, D. B. Oerther, Bifidobacteria in feces and environmental waters. *Appl. Environ. Microbiol.* **74**, 575-584 (2008).
488. L. Rouli, V. Merhej, P. E. Fournier, D. Raoult, The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect* **7**, 72-85 (2015).