Alma Mater Studiorum – Università di Bologna
in cotutela con Graduate University for Advanced Studies, SOKENDAI, Japan

DOTTORATO DI RICERCA IN
SCIENZE E TECNOLOGIE AGRARIE, AMBIENTALI E ALIMENTARI

Ciclo 36

Settore Concorsuale: 07/I1

Settore Scientifico Disciplinare: AGR/16

Comparative Analysis of Genus *Bifidobacterium*: Insight into its Host Adaptation

Presentata da: Maria Altaf Satti

Coordinatore Dottorato

Massimiliano Petracci

Supervisore

Masanori Arita

Supervisore

Paola Mattarelli

Esame finale anno 2021

Declaration

I hereby declare that this thesis comprises of my own research work and on bases of efforts made under the sincere guidance of my supervisor **Prof. Masanori Arita**. None of the part of this thesis is plagiarized. The contributions of different people in this work is acknowledged and duly referred.

Maria Altaf Satti

“What we know is a drop, what we don’t know is an ocean.”

Isaac Newton

Dedicated to my Parents

(Altaf Ahmed Satti & Nasreen Akhter)

Acknowledgments

I would like to express my deepest appreciation to all those who helped me and guided me through my PhD journey. Today I am able to complete my dissertation because of the guidance and support of these people.

First of all, I want to express my wholehearted thanks to my supervisor Prof. Masanori Arita for his continuous support and expert guidance. Without his fortitude, inspiration, encouragement, effort and counsel my thesis would not have been completed. His immense knowledge and guidance always helped me whenever I was stuck in my research. His keen interest and support encouraged me to put in my best efforts. His kindness and concern make my stay in Japan easy and enjoyable. I will always be indebted to him for the support and encouragement he has given to me.

Next, I would like to pay immense thanks to Prof. Paola Matterelli from University of Bologna, who served as my sub supervisor and gave insightful suggestions and guidance for my work. It's because of her efforts that I would be able to apply for cotutelle and get a dual PhD degree from University of Bologna. I am indebted to her for her assistance in numerous ways from providing me with the opportunity to present my work in workshop and conferences to making my stay in Italy enjoyable. I would also like to thank all the members in her laboratory specially Monica Modesto for her thoughtful suggestions.

I would like to thank my guidance and examination committee members: Profs. Niki Hironori, Ikeo Kazuho, Nakamura Yasukazu and Miyagishima Shin-ya, for their insightful comments, encouragement and guidance. Their comments always encouraged me to do better. I am also thankful to Prof. Akihito Endo, Assistant Prof. Tanizawa Yasuhiro and Assistant Prof. Kawashima Takeshi for their suggestions and feedback.

I am grateful to all my lab members for their useful discussion and providing a friendly

environment. I would like to pay special thanks to the lab secretaries Ohnuki-san and Murakata-san for helping me settle in and were welcoming to solve and help out every small matter.

I am greatly thankful to SOKENDAI for providing me the opportunity for international collaboration. With the funds provided under “SOKENDAI Student Dispatch Program” I was able to visit laboratory of Prof. Paola Matterelli in University of Bologna twice in years 2018 and 2019. This visit resulted in a fruitful and strong collaboration in terms of scientific knowledge sharing and research publications.

My sincere gratitude to the Ministry of Education, Sports, Culture, Science and Technology (MEXT), Japan for providing me the opportunity to pursue my doctoral degree in Japan.

I want to express my sincere thanks to my friend Mehwish Noureen for her support and kind suggestions. Last but not the least, special thanks to my parents, brother, sisters and grandmother for their love, moral support and motivation.

Abstract

Human gut is a home to the trillions of microorganisms, and many of the bacterial genera are known as probiotics. Bifidobacteria is amongst one of these health promoting bacteria imparting beneficial health effects on their host by immunomodulation and metabolic activities. The genus *Bifidobacterium* is a ubiquitous, probiotic group in the phylum Actinobacteria, and exist in anaerobic gut environments of various host species, from insects like bees to mammals. The role of these important probiotic genera can be elucidated by understanding their genomes. Comparative analysis of the whole genus of these bacteria can reveal their adaptation to a diverse host range.

This study comprises of four research projects. The first focuses on providing the accurate annotations and selection of the core genome. The second aims to explore the interaction of bifidobacteria with its host by investigating their extracellular structures. The last two focus on the adaptation of bifidobacteria to diverse host range and elucidate the relationship of bifidobacteria with their host diet.

In the first study, a public library of gene functions in the genus *Bifidobacterium* for its online annotation was prepared. The core genes in each genus were selected based on a newly proposed statistical definition of core genome. Comparative analysis of genus *Bifidobacterium* with another probiotic genus *Lactobacillus* revealed the metabolic characteristics of genus *Bifidobacterium*. The analysis showed that the protein families overrepresented in *Bifidobacterium* were mostly involved in complex sugar metabolism host interaction, and stress responses.

The second study investigated the immunomodulatory role of *B. bifidum* strain TMC3115, isolated from healthy infant. *B. bifidum* TMC3115 is an important strain isolated from healthy infant. This strain exhibits inhibitory effect in allergic inflammation. The analysis of TMC3115 provided insights into its extracellular structures which might have their role in host interaction and immunomodulation. The study highlighted the variability among the *Bifidobacterium* genomes just not on species level but also on strain level in terms of host interaction.

The last two studies aim to inspect the relationship between bifidobacteria and its host diet. Bifidobacteria, being an obligatory anaerobic species, are both host- and niche-specific. The genetic biodiversity was investigated for bifidobacteria from bat, human and to non-human primates. The investigation of bifidobacterial species from different niches or hosts is fundamental in clarifying the repertoire of genes that have caused their evolutionary differentiation. Such adaptation of bifidobacterial species is considered relevant to the intestinal microecosystem and hosts' oligosaccharides including those of food and milk. Many species should have co-evolved with their hosts, but the phylogeny of *Bifidobacterium* is dissimilar to that of host animals. The discrepancy could be linked to the niche-specific evolution due to hosts' dietary carbohydrates. Since carbohydrates are the main class of nutrients for bifidobacterial growth, the distribution of carbohydrate-active enzymes, in particular glycoside hydrolases (GHs) that metabolize unique oligosaccharides was examined. When bifidobacterial species were classified by their distribution of GH genes, five groups arose according to their hosts' feeding behavior. The distribution of GH genes was only weakly associated with the phylogeny of the host animals or with genomic features such

as genome size. Thus, the hosts' dietary pattern is the key determinant of the distribution and evolution of GH genes.

List of Publications

1. **Satti M**, Tanizawa Y, Endo A, Arita M. Comparative analysis of probiotic bacteria based on a new definition of core genome. *Journal of bioinformatics and computational biology*.2018;16(03):1840012.
2. Modesto M, Watanabe K, Arita M, **Satti M**, Oki K, Sciavilla P, et al. *Bifidobacterium* jacchi sp. nov., isolated from the faeces of a baby common marmoset (Callithrix jacchus). *International journal of systematic and evolutionary microbiology*. 2019;69(8):2477-2485.
3. Modesto M*, **Satti M***, Watanabe K, Puglisi E, Morelli L, Huang CH, et al. Characterization of *Bifidobacterium* species in feaces of the Egyptian fruit bat: Description of *B. vespertilionis* sp. nov. and *B. rousetti* sp. nov. *Systematic and applied microbiology*. 2019;42(6):126017.
4. Modesto M, **Satti M**, Watanabe K, Sciavilla P, Felis GE, Sandri C, et al. Alloscardovia theropitheci sp. nov., isolated from the faeces of gelada baboon, the'bleeding heart'monkey (Theropithecus gelada). *International journal of systematic and evolutionary microbiology*. 2019;69(10):3041-3048.
5. Modesto M*, **Satti M***, Watanabe K, Scarafile D, Huang CH, Liou JS, et al. Phylogenetic characterization of two novel species of the genus *Bifidobacterium*: *Bifidobacterium saimiriisciurei* sp. nov. and *Bifidobacterium platyrrhinorum* sp. nov. *Systematic and Applied Microbiology*.2020;43(5):26111.
6. Modesto M, **Satti M**, Watanabe K, Huang CH, Liou JS, Tamura T, et al. Bifidobacteria in two-toed sloths (*Choloepus didactylus*): phylogenetic characterization of the novel taxon *Bifidobacterium choloepi* sp. nov. *International Journal of Systematic and Evolutionary Microbiology*. 2020;70(12):6115-6125.
7. **Satti M**, Modesto M, Endo A, Kawashima T, Mattarelli P, Arita M. Host-Diet Effect on the Metabolism of *Bifidobacterium*. *Genes*. 2021;*12*(4):609.

***co-first authors**

Table of Contents

List of Figures

List of Tables

CHAPTER 1

General Introduction

The gastrointestinal tract (GIT) of mammals is colonized by diverse range of bacteria known as gut microbiota. In humans there are trillions of these bacteria [1]. Among them many of the bacterial genera impart beneficial effect on host health known as probiotics. They are available in the market and claim to impart healthy benefits on consumer health by their interaction with the human GIT. Some probiotic species by their competition for attachment sites, production of antimicrobials and modulating host-acquired immune system prevent other pathogens from colonizing the intestine [2]. Bifidobacteria is amongst one of these health promoting bacteria residing in gut. It is a probiotic bacterium and most members of the genus are considered safe and non-toxic to human. This genus is considered to be one of the first colonizer of the neonates [3]. They were dominantly found in the gut of the infants [4].

1.1. Discovery of *Bifidobacterium*

The phylum Actinobacteria is among the most diverse and abundant group of microorganisms in nature [5,6]. They are gram positive bacteria with high G+C content ranging from 51 to 77 percent. The bacteria in this phylum have the ability to produce natural bioactive compounds and are adapted to diverse ecosystems. The phylum includes pathogens (e.g., *Mycobacterium* spp., *Corynebacterium* spp.) as well as gut commensals (*Bifidobacterium* spp.).

The genus *Bifidobacterium* belongs to family Bifidobacteriaceae and was first isolated in 1899 by Henri Tissier at the Pasteur Institute in Paris from feces of healthy breast-fed infants [7]. It was initially named as *Bacillus bifidus* (meaning ‘divided into two parts’ in Latin for its Y-shape), and later reclassified as *Lactobacillus bifidus*. In 1924, Orla-Jensen reclassified it into an independent genus *Bifidobacterium*. Bifidobacteria represent the Gram-positive, immotile, non-gas-producing, obligatory anaerobic, Y-shaped bacteria, and possess a high G+C genome content [7,8].

1.2. Bifidoacterial ecology

Bifidobacteria inhabit in a wide variety of hosts including mammals, birds and insects [9-14]. Certain species of bifidobacteria can be found in environmental niches like sewage, fermented

products, and anaerobic digesters [15-17]. Bifidobacteria varies in host specificity, examples of host-specific species are *Bifidobacterium breve* for humans, *Bifidobacterium rousetti* for bat and *Bifidobacterium reuteri* for marmoset. On the other hand, there are some species with cosmopolitan life style such as *Bifidobacterium longum*, isolated from humans and animals, and *Bifidobacterium animalis* and *Bifidobacterium pseudolongum* isolated from different animal species [18].

In human, they are present in GIT and oral cavity. Notably they represent the dominant clade of the gut microbiota of healthy, breast-fed infants. For this reason, the commensal species are considered important for microbial modulation at birth, such as the immune programming of its host [19,20]. In infants' vertical transmission and subsequent breastfeeding results in the development of bifidobacterial species [21,22]. Several species of bifidobacteria undergoes genetic and metabolic adaptations to colonize the gut. For instance, the species more prevalent in infants (*B. breve, Bifidobacterium longum* subsp. *infantis, Bifidobacterium longum* subsp. *longum, Bifidobacterium pseudocatenulatum,* and *B. bifidum*) [23], have ability to metabolize certain oligosaccharides present in human milk whereas the species *Bifidobacterium adolescentis, Bifidobacterium catenulatum, B. pseudocatenulatum,* and *B. longum subsp. longum* are commonly found in adults can metabolize complex plant derived carbohydrates [24,25]. There is a great diversity in bifidobacterial species and strains even among the same host. Bifidobacterial distribution changes within the different ages in human, however not much is known about the diversity of bifidobacterial species among the various compartments of GIT of same individual [26].

1.3. Useful features of bifidobacteria

Bifidobacteria are saccharolytic organisms, they encompass a wide range of enzyme encoding genes involved in the uptake and catabolism of complex and non-digestible carbohydrates including those from milk oligosaccharides to plant fibers [27]. Bifidobacteria possess a unique metabolic pathway known as "bifid shunt", which degrade the hexose sugars glucose and fructose with the key enzyme fructose-6-phosphate phosphoketolase [28]. This ATP generating pathway mostly generates short chain fatty acids (SCFAs) which antagonise pathogenic bacteria [29]. For example, the acetate produced by bifidobacteria protects the host against the pathogenic infections [30].

Bifidobacterial species has a potential to treat various gastrointestinal disorders such as diarrhea, necrotizing enterocolitis and inflammatory bowel disease [31-33]. The species of

B. bifidum, *B. breve* and *B. longum* are mostly used to treat these disorders. Bifidobacteria have also been reported to prevent gastrointestinal disorders by competitive exclusion of pathogenic bacteria. For instance, *Clostridium perfringens*, a known producer of undesirable toxins was reduced by the presence of high number of *Bifidobacterium* [34]. Some strains of Bifidobacteria like *B. animalis* BF052, and *Bifidobacterium animalis* subsp. *lactis* BB-12 are used as an active ingredient in many commercial probiotic products [35]. The probiotic activity of *Bifidobacterium* is strain dependent [36].

Bifidobacteria also plays role in immunostimulation, pathogen exclusion and vitamin production. They can produce folate (B-vitamin), which is considered important nutrient for cell metabolism and immune development [37]. The ability of bifidobacteria to produce folate is strain dependent. Non-human resident bifidobacteria produce relatively lower amount of folate compared to human resident [38]. As bifidobacteria exert such beneficial effects they are widely utilized in food industry.

1.4. Genomes of bifidobacteria

Currently the genus *Bifidobacterium* encompasses 88 recognized taxa with 80 species and 8 subspecies (*B. animalis* subsp. *lactis*, *B. longum* subsp. *infantis*, *Bifidobacterium longum* subsp. *suis*, *Bifidobacterium catenulatum* subsp. *kashiwanohense*, *Bifidobacterium pseudolongum* subsp. *globosum*, *Bifidobacterium pullorum* subsp. *gallinarum*, *Bifidobacterium pullorum* subsp. *saeculare*, *Bifidobacterium thermacidophilum* subsp. *thermacidophilum*). Seventy-nine taxa were isolated from the gastrointestinal tract of mammals, birds, or insects and nine were isolated from sewage or fermented milk [39]. The first sequenced complete genome in *Bifidobacterium* genus was for *B. longum* subsp. *longum* NCC2705 [40]. Since then new genomes of *Bifidobacterium* from the species isolated from various sources were sequenced. Among the 88 taxa; 18 are complete genomes and 70 are draft genomes. Table 1.1 shows the list of the 88 recognized *Bifidobacterium* taxa.

The average genome size of the *Bifidobacterium* species is 2.2 Mb with the smallest genome of *Bifidobacterium commune*,1.6 Mb and largest of *Bifidobacterium biavatii*, 3.3 Mb. The G+C content ranges from 53% to 66% and the number of coding sequences (CDS) ranged from 1200 to 2500 [41].

Table 1.1. List of *Bifidobacterium* (sub)species recognized as type strains.

Species	Specific host Information
B. actinocoloniiforme DSM 22766	Bumble bee digestive tract
B. adolescentis ATCC 15703	Intestine of human Adult
B. aemilianum LMG 30143	Carpenter bee digestive tract
B. aerophilum DSM 100689	Feces of cotton-top tamarin
B. aesculapii DSM 26737	Feces of common marmosets
B. angulatum DSM 20098	Feces of human
B. animalis subsp. *animalis* ATCC 25527	Feces of mouse
B. animalis subsp. *lactis* DSM 10140	Commercial probiotic
B. anseris LMG 30189	Feces of domestic goose
B. apri DSM 100238	Feces of wild pig
B. aquikefiry LMG 28769	Water kefir
B. asteroides DSM 20089	Hindgut of honey bee
B. avesanii DSM 100685	Feces of cotton-top tamarin
B. biavatii DSM 23969	Feces of red-handed tamarind
B. bifidum ATCC 29521	Infant stool
B. bohemicum DSM 22767	Bumble bee digestive tract
B. bombi DSM 19703	Bumble bee digestive tract
B. boum DSM 20432	Feces of cattle
B. breve DSM 20213	Infant stool
B. callimiconis LMG 30938	Feces of goeldi's marmoset
B. callitrichidarum DSM 103152	Feces of emperor tamarin
B. callitrichos DSM 23973	Feces of common marmosets
B. canis DSM 105923	Feces of dog
B. castoris LMG 30937	Feces of European beaver
B. catenulatum subsp. *catenulatum* LMG 11043	Infant stool
B. catenulatum subsp. *kashiwanohense* DSM 21854	Infant stool
B. catulorum DSM 103154	Feces of common marmosets
B. cebidarum LMG 31469	Feces of golden-headed tamarin
B. choerinum LMG 10510	Piglet feces
B. cholopei BRDM6	Feces of sloth
B. commune DSM 28792	Bumble bee digestive tract
B. coryneforme LMG 18911	Hindgut of honey bee

Table 1.1. (Continued)

B. criceti LMG 30188	Feces of European hamster
B. crudilactis LMG 23609	Raw cow milk
B. cuniculi LMG 10738	Feces of rabbit
B. dentium LMG 20436	Oral cavity
B. dolichotidis LMG 30941	Feces of Patagonian mara
B. eulemuris DSM 100216	Feces of adult black lemurs
B. felsineum DSM 103139	Feces of cotton-top tamarin
B. gallicum LMG 11596	Adult intestine
B. goeldii LMG 30939	Feces of goeldi's marmoset
B. hapali DSM 100202	Feces of common marmosets
B. imperatoris LMG 30297	Feces of emperor tamarin
B. indicum LMG 11587	Hindgut of honey bee
B. italicum LMG 30187	Feces of rabbit
B. jachii DSM 103362	Feces of common marmosets
B. lemurum DSM 28807	Feces of ring-tailed lemur
B. leontopitechi LMG 31471	Feces of Goeldi’s monkey
B. longum subsp. *infantis* DSM 20088	Intestine of inafant
B. longum subsp. *longum* JCM 1217	Intestine of adult
B. longum subsp. *suis* DSM 20211	Feces of pig
B. magnum LMG 11591	Feces of rabbit
B. margollesii LMG 30296	Feces of pygmy marmoset
B. merycicum LMG 11341	Feces of cattle
B. minimum LMG 11592	Sewage
B. mongoliense DSM 21395	Fermented mare’s milk
B. moukalabense DSM 27321	Feces of wild lowland gorilla
B. myosotis DSM 100196	Feces of common marmosets
B. parmae LMG 30295	Feces of pygmy marmoset
B. platyrrhinorum DSM 106029	Feces of squirrel monkey
B. primatium DSM 100687	Feces of cotton-top tamarin
B. pseudocatenulatum LMG 10505	Infant feces
B. pseudolongum subsp. *globosum* DSM 20092	Rumen of bovine
B. pseudolongum subsp. *pseudolongum* LMG 11571	Feces of pig
B. psycraerophilum LMG 21775	Fermented product
B. pullorum subsp. *gallinarum* LMG 11586	Chicken cecum
B. pullorum subsp. *pullorum* DSM 20433	Feces of chicken
B. pullorum subsp. *saeculare* LMG 14934	Feces of rabbit

Table 1.1. (Continued)

B. ramosum DSM 100688	Feces of cotton-top tamarin
B. reuteri DSM 23975	Feces of common marmosets
B. rousetti DSM 106027	Feces of Egyptian fruit-bat
B. ruminantium LMG 21811	Feces of cattle
B. saguini DSM 23967	Feces of red-handed tamarind
B. samirii LMG 30940	Feces of black-capped squirrel monkey
B. saimiriisciurei DSM 106020	Feces of squirrel monkey
B. scaligerum DSM 103140	Feces of cotton-top tamarin
B. scardovii DSM 13734	Blood
B. simiarum DSM 103153	Feces of emperor tamarin
B. stellenboschense DSM 23968	Feces of red-handed tamarind
B. subtile LMG 11597	Sewage
B. thermacidophilum subsp. *porcinum* DSM 17755	Feces of Piglet
B. thermacidophilum subsp. *thermacidophilum* LMG 15837	Sewage
B. thermophilum JCM 1207	Feces of adult
B. tibiigranuli LMG 31086	Water kefir
B. tissieri DSM 100201	Feces common marmosets
B. tsurumiense JCM 13495	Hamster dental plaque
B. vansinderenii LMG 30126	Feces of emperor tamarin
B. vespertilionis DSM 106025	Feces of Egyptian fruit-bat
B. xylocopae LMG 30142	Carpenter bee digestive tract

1.5. Bifidobacterial phylogeny

In recent years *Bifidobacterium* phylogeny is mostly characterized based on different methods: 16S rRNA gene, by considering multilocus housekeeping genes (i.e. *hsp60, clpC, dnaJ, dnaG* and *rpoB*) [42,43] and core genes-based tree. Studies show that the tree based on core genes is more robust than the 16S rRNA based gene tree [44,45]. A comparative genomic analysis based on 84 type strains represent 362 core genes. The phylogenetic tree based on these 362 core genes revealed 11 different phylogenetic groups: *B. adolescentis, B. boum, B. pullorum, B. asteroides, B. longum, B. psychraerophilum, B. bifidum, B. pseudolongum, B. bombi, B. tissieri* and *B. vespertilionis* group (Figure 1.1) [46].

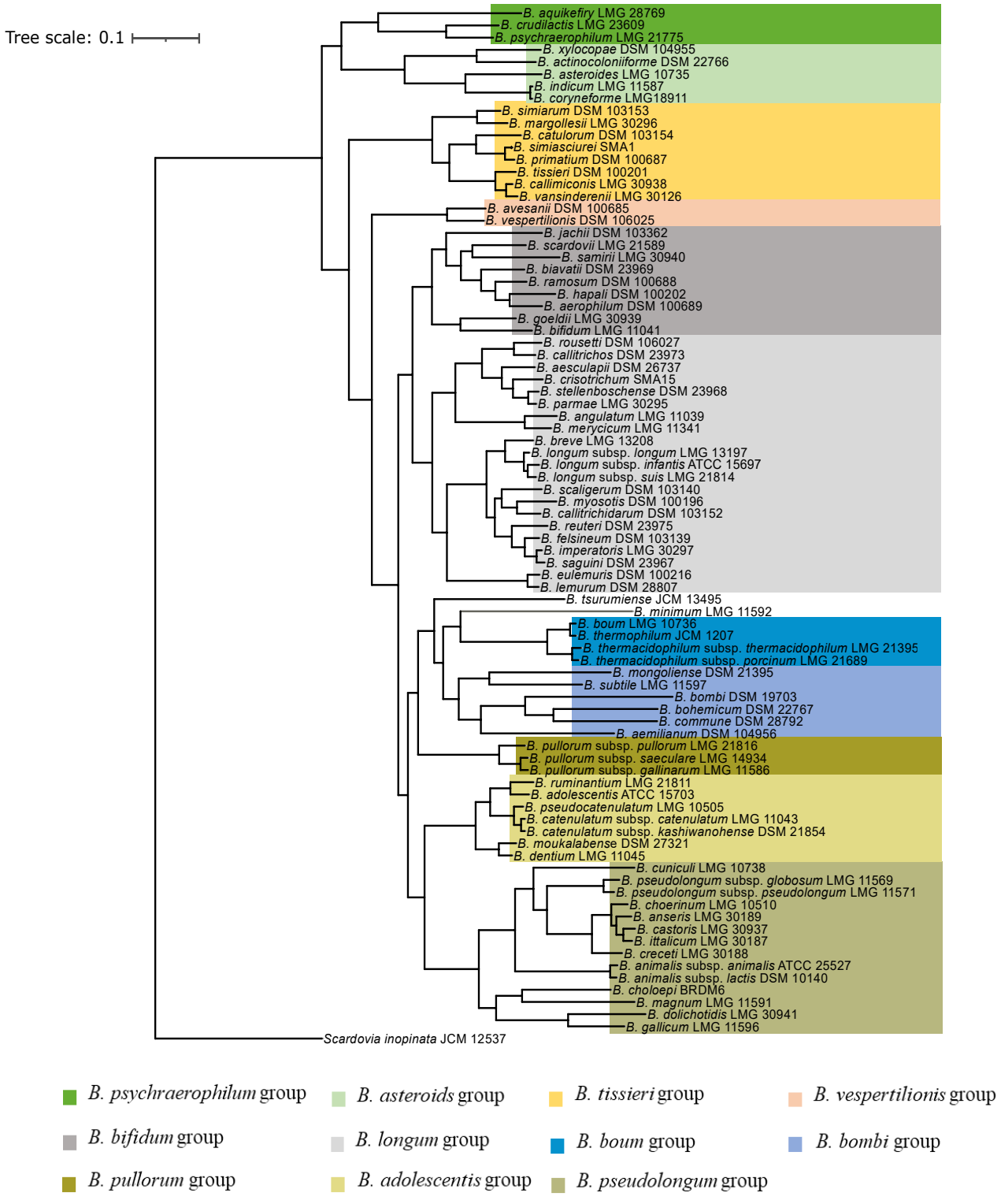


Figure 1.1. Phylogenetic tree based on concatenated amino acid sequences of 362 core genes of the 84 type strains. Eleven phylogenetic groups are highlighted in different colors.

1.6. Bifidobacteria and host interaction

To interact with their host the commensal bacteria have evolved through specific strategies. Although *Bifidobacterium* has wide range of health benefits, however the mechanisms that how they interact with their host is yet not clear. The various extracellular structures including those of pili, capsular polysaccharides or exopolysaccharides and some bioactive metabolites seems to play important role in host interaction and thus modulating the immune system [47,48].

The bifidobacterial genomes encode two different types of pili structures, i.e. sortase dependent pili and type IV or tight adherence pili (Tad pili). The comparative analysis of the *Bifidobacterium* genomes has shown that there is diversity among the species and strains in terms of the number and sequence variability for the sortase dependent pili. For example, the number of these sortase dependent pilus loci varies from the total absence of these in some *Bifidobacterium* species like *Bifidobacterium actinocoloniforme* and *B. longum* to the strain of *Bifidobacterium dentium* Bd1 having upto seven of these pili encoding loci [49,50]. The detailed study of sortase dependent pili in *B. bifidum* PRL2010 revealed that these pili were able to induce high levels of TNF-α cytokines and also reduce expression of other proinflammatory cytokines. This suggest the role of these pili structures in immune modulation [51]. Like the sortase dependent pili, Tad pili also plays role in host interaction. The gene cluster encoding for Tad pilus is highly conserved in bifidobacterial genomes [52]. They might contribute to maturation of epithelial cells of new-borns and thus maintaining the host mucosal homeostasis [53].

Other key extracellular structure involved in bifidobacterial host interaction are exopolysaccharides (EPSs). Many studies have revealed the role of bifidobacterial EPS in gut colonization [54-56]. A comparative analysis of 48 bifidobacterial species shows that all the *Bifidobacterium* species have at least one EPS cluster except the *B. bifidum* species [57]. The EPS structures also play important role in immune modulation. For instance, the strain of *B. breve* UCC2003 produces EPS which has ability to modulate immune response and reduce the gut pathogen infection [58].

In addition to pili and EPS, bifidobacterial serine protease inhibitor (serpin) is also involved in host-microbe interaction. Genome analysis of *Bifidobacterium* species revealed that serpin-like gene is not ubiquitous and present in some species, such as *B. breve*, *B. longum* subsp. *longum*, *B. longum* subsp. *infantis*, *B. longum* subsp. *suis*, *Bifidobacterium cuniculi*, *Bifidobacterium scardovii*, and *B. dentium* [59].

1.7. Carbohydrate metabolism in bifidobacteria

One of the key mechanisms for colonization and survival of bacteria in the gastrointestinal tract is its ability to degrade complex carbohydrates. These complex carbohydrates are either host derived compounds (e.g., mucin and human milk oligosaccharides) or dietary compounds (e.g., xylan, starch, cellulose, hemicellulose, pectin and gums) [60,61]. These complex carbohydrates pass to the lower gut as they are not metabolized by the host and microbes in upper gut. Here these indigestible carbohydrates are metabolized by certain gut commensals including the species of genus *Bifidobacterium*. The genomes of bifidobacteria have a large number of genes encoding the enzymes involved in metabolism of complex carbohydrates. For instance, the genomes of *B. bifidum* PRL2010 and *B. longum* subsp. *infantis* ATCC15697 have the enzymes encoding for host glycan degradation and *B. adolescentis* species can utilize certain dietary carbohydrates such as resistance starches [60-62].

Among the bifidobacterial strains their ability to metabolize carbohydrates differs considerably. Many of the characterized strains can utilize ribose, galactose, fructose, glucose, sucrose, maltose, melibiose and raffinose, but cannot degrade L-arabinose, rhamnose, *N*-acetylglucosamine, sorbitol and trehalose [28].

1.7.1. *Carbohydrate import in bifidobacteria*

Bifidobacteria have the ability to metabolize range of mono-, di- and oligo-saccharides which are mainly transported into their cytoplasm by ATP-binding cassette transporters (ABC-type transporters), major facilitator superfamily (MFS) transport system and Phosphoenol Pyruvate-Phosphotransferase Systems (PEP-PTSs) [63,64]. Among these transport systems, ABC transporters are most common in bifidobacteria. For example, *B. longum* subsp. *longum* NCC2705 and *B. longum* subsp. *infantis* ATCC15697 possesses around 13 ABC transporters and less than 3 of the other transport systems [65,60]. However, there are exceptions; *B. bifidum* PRL2010 genome preferentially uses PEP-PTSs system for carbohydrate utilization as it encodes four PEP-PTSs systems and two ABC transporters [66].

1.7.2. *Enzymatic degradation of carbohydrates by bifidobacteria*

The genes encoding for carbohydrate-active enzymes (CAZymes) are of special interest in gut microbiome, as these enzymes are required to digest complex dietary polysaccharides. CAZymes encoded by gut microbiome catalyze the breakdown of oligosaccharides and

polysaccharides to fermentable monosaccharides. There are two types of enzymes that breakdown the glyosidic bond between two or more monosaccharides or between a carbohydrate and non-carbohydrate moiety: Glycosyl hydrolases (GHs) and Polysaccharide lyases (PLs). GHs breakdown the bonds by the insertion of water molecule (hydrolysis) and PLs breakdown the complex carbohydrates by the elimination mechanism [67, 68]. The classification of CAZymes in families based on amino acid similarity is available at Carbohydrate-Active Enzyme (CAZy) database (http://www.cazy.org/). This database provides the classification of enzymes involved in synthesis and metabolism of complex carbohydrates [67]. Other than GHs and PLs CAZy database lists two additional CAZymes categories, carbohydrate esterases (CEs), which facilitate the action of GHs and PLs by removing ester substituent from glycan chains, and glycosyl transferases (GTs), which assemble complex carbohydrate from activated sugar donors [67]. The metabolism of complex carbohydrates such as plant pectin requires multiple enzymes, which are encoded usually in a multigene locus known as polysaccharide utilization loci (PULs). Previous studies reported that most genes encoding GHs formed PULs along with genes encoding transporters and regulators [69].

In bifidobacteria, GHs are the most important group of enzymes which help them to adapt to the diverse environment by hydrolysis of complex diet-derived and host produced carbohydrates. The genes encoding for enzymes involved in metabolism of carbohydrates account for 13.5% in the pan-genome and 5.5% in the core genome of *Bifidobacterium* [70]. Bifidobacteria possess one of the largest arsenals of GH13 (α-1,4-glucosidase, amylopullulanase, sucrose phosphorylase, α-amylase), GH43(Endo-1,4-β-xylanase, β-1,4-xylosidase), GH3(β-glucosidase, β-hexosaminidase, β-glucosideglucohydrolase) and GH51 (α-L-arabinofuranosidase) among various gut bacteria [71]. The most abundant GHs family in *Bifidobacterium* is GH13. The enzymes in this family are involved in the hydrolysis of complex carbohydrates such as glycogen, starch, amylose, amylopectin, pullan, maltodextrin as well as glycans present in adult mammalian diet like stachyose raffinose and melibiose [72]. Bifidobacteria possess GH families for the degradation of host glycan. For instance, they have GH involved in metabolism of milk carbohydrates i.e. GH33, GH34 (exo-sialidases), GH29 and GH30 (fucosidases) and GH20 (hexosaminidase and lacto-*N*-biosidase) [71].

1.8. Organization and purpose of the dissertation

Bifidobacteria is one of the dominant bacterial group in the GIT of human and other animals and have beneficial effects on host health. Bifidobacteria has a diverse host range and are generally considered as host-animal specific bacteria. The role of these important probiotic genera can be elucidated by understanding their genomics. Comparative analysis of the whole genus of these bacteria can be very helful to completely understand their adaptation to a diverse host range.

This study aims to investigate the interaction and adaptation of bifidobacteria with their host by using a comparative genomic approach. The relationship between bifidobacterial species and their host is still not clear. There is discrepancy in relationship of host environment and diversity of *Bifidobacterium.* However, the extracellular structures involved in intestinal epithelial adhesion or metabolization of host or diet derived compounds are thought to have some role in this relationship. There is abundant knowledge that how bifidobacterial species are related with their human host but very less is known about other hosts specifically the non-human primates.

In this study we focus on exploring this relationship of bifidobacterial species with their host by genomic analysis, considering all the sequenced type strains of *Bifidobacterium*. More specifically the studies in this dissertation focuses on investigating the genus *Bifidobacterium* aiming to inspect the genetic adaptation of bifidobacterial species to a diverse host range and to examine the evolutionary relationship between bifidobacteria and host animals.

This dissertation consists of five main chapters. The second focusing on providing the accurate annotations and selection of the core genome. The third aims to explore the interaction of bifidobacteria with its host by investigating their extracellular structures. The last two focuses on the adaptation of bifidobacteria to diverse host range and elucidate the relationship of bifidobacteria with their host diet.

In Chapter 2, the creation of reference library and comparative analysis of genus *Bifidobacterium* and *Lactobacillus* based on new proposed definition of core genome is described. A public library of gene functions in the genus *Bifidobacterium* for its online annotation was prepared. The core genes in each genus were selected based on a newly proposed statistical definition of core genome: for *Bifidobacterium* gene clusters present in at

least 92% of genomes and for *Lactobacillus*, 97% makes their core genome. The functional comparison of core and pan genomes showed that there is little difference in their pan genomes but a significant difference in their core genomes, specifically within the "amino acid transport and metabolism" and "translation, ribosomal structure and biogenesis" categories. Overrepresented *Bifidobacterium* protein families were mostly involved in host interaction, complex compounds metabolism and stress responses. The reference library for genus *Bifidobacterium* enabled the accurate and consistent annotations. Based on a statistical analysis of pan and core genomes, the study revealed the metabolic difference between two genera and investigated over- and underrepresented protein families in *Bifidobacterium* relative to *Lactobacillus*. The differential study could reveal host interaction and adaptability in *Bifidobacterium*, together with broad adaptability for amino acids and carbohydrate metabolism.

In Chapter 3, The detailed genomic structure and the genomic features of the *B. bifidum* strain TMC3115 and its role in host interaction and immunomodulation is described. *B. bifidum* species are among the first colonizer of gastrointestinal tract of the neonates. *B. bifidum* TMC3115 is an important strain isolated from healthy infant. This strain exhibits inhibitory effect in allergic inflammation. This study aims to explore the genome structure, features and the immunomodulatory role of this strain. The genomic analysis showed that the genome of TMC3115 strain have an inversion of ~ 382 kb. Although the inversion disrupts the replication symmetry yet the inversion affects the growth and genomic integrity is not definite. The strain possesses important extracellular proteins with binding domains involved in host interaction. The sortase dependent pili (SD pili) of TMC3115 shows high homology with SD pili of PRL2010 strain where these pili were found to be involved in immunomodulatory activity. The comparative analysis of SD pili showed that there is diversity among the *B. bifidum* strains in the number and sequence of pili.

Chapter 4 describes the genetic biodiversity of bifidobacteria from bat compared to bifidobacterial species from human and non-human primates. The comparative analysis in this study has revealed the important features of bifidobacteria in bat such as their contribution in metabolizing the host dietary carbohydrates. Bat and non-human primate specific GHs corresponding to the metabolism of their dietary carbohydrates suggest the dietary association between these groups. The description of the genomic features in different niches (bat, non-human primates and human being) is fundamental in clarifying repertoire of genes that have caused their evolutionary differentiation. Such genomic analyses support the hypothesis that

bat strains have been subjected to genetic adaptations to their host environment such as a peculiar diet heavily based on sugars.

In chapter 5, the relationship between bifidobacteria and their host diet using a comparative genomics approach is described. *Bifidobacterium* has a diverse host range and shows several beneficial properties to the hosts. Many species should have co-evolved with their hosts, but the phylogeny of *Bifidobacterium* is dissimilar to that of host animals. The discrepancy could be linked to the niche-specific evolution due to hosts' dietary carbohydrates. Since carbohydrates are the main class of nutrients for bifidobacterial growth, the distribution of carbohydrate-active enzymes, in particular glycoside hydrolases (GHs) that metabolize unique oligosaccharides were examined. When bifidobacterial species are classified by their distribution of GH genes, five groups arose according to their hosts' feeding behaviour. The distribution of GH genes was only weakly associated with the phylogeny of the host animals or with genomic features such as genome size. Thus, the hosts' dietary pattern is the key determinant of the distribution and evolution of GH genes.

CHAPTER 2

Bifdobacterium Reference Library and Comparative Analysis based on a New Definition of Core Genome

2.1. Introduction

The genus *Bifidobacterium* and *Lactobacillus* has probiotic properties due to which they are widely used in the food industry. These microorganisms are commensal and considered as imparting health improving benefits on their hosts. The species in these genera are considered as safe and does not cause diseases, however they produce important compounds such as lactic acid, antimicrobials and bacteriocins [73]. They play beneficial role in their host like strengthen the immune system, prevent different diseases and protect against harmful microorganisms [8].

The role of these important probiotic genera can be elucidated by understanding their genomics. Nowadays with the availability of large sequencing data for these probiotic bacteria deep insights into molecular mechanisms, their interaction with the host and their genetic basis for imparting health improving effects can be made. Comparative analysis of the whole genus of these bacteria can be done to completely understand the mechanisms by which these probiotic bacteria impart beneficial impacts on its host.

For genomic analysis, it is essential to have an accurate genome annotation for the genomes under analysis. With the availability of large number of sequencing data, it is hard to do experimental validations for the annotations. We can use the computational pipelines to annotate the genomes but it is important to have an accurate reference library against which the homology search is done to assign the gene functions. Often these libraries are more generalized and not accurate. The approach to use the specified reference library which is manually curated is important for having the accurate genome annotations.

In bacterial genomic studies defining the core genome is often the first step. The traditional definition of selecting the core genome is the number of genes present in 100% of the genomes under analysis; however, this approach has some problems. If the dataset of interest has more diversity among their genomes than the core genome would be smaller in

comparison to the dataset where genomes have less diversity. More generally the number of the core genes is highly dependent on the size of dataset [73].

This study focused on providing the accurate and consistent genomic annotation for this beneficial genus and compared its probiotic characteristics with similar commensal probiotic genus, *Lactobacillus*. For this assessment, I constructed a reference annotation library, which is freely available at DFAST annotation server, and provided a statistical definition for the "core genome." Based on this definition, genus-specific metabolic functions for *Bifidobacterium* and *Lactobacillus* could be identified.

2.2. Method

2.2.1. *Construction of the Bifidobacterium gene library*

For a consistent genome annotation, a reference library is required (pairs of sequence and its functional annotation) that is compatible with published complete genome annotations. To prepare the library, complete genomes of 67 strains from *Bifidobacterium*, 8 strains from *Lactobacillus* and 1 strain from *Bacillus subtilis* were collected from the NCBI Assembly Database. The protein sequences for all the genomes were extracted and subjected to orthologous clustering. Orthologous clusters were generated using GET_HOMOLOGUES software [74]. The parameters for the clustering were as follows. The E-value threshold was 10e-5, the minimum percentage coverage was 75%, and the clustering algorithm was OrthoMCL [75]. A total of 144,028 identified protein sequences were grouped into 21,255 clusters, among which 12,545 were singletons. The singletons were discarded and the remaining 8,710 clusters were further analyzed. For their protein names, gene symbols and EC numbers, our annotation library for *Lactobacillus* was first referenced [76]. Among 8,710 clusters, functions of 6,697 clusters were identified, indicating that close to 80% of the shared genes in *Bifidobacterium* have orthologs in *Lactobacillus*. Functions of remaining gene clusters were manually sought by referencing the NCBI Conserved Domain Database (NCBI-CDD) using the Reverse Position-Specific BLAST. Through this process, 15 *Bifidobacterium* strains were found as close duplicates. In the following analysis, I excluded them and used the remaining 52 strains. For 45,038 gene clusters in 178 *Lactobacillus* genomes, the annotation results at DFAST web server were used [76]. A newly sequenced *Bifidobacterium* can be easily annotated through this system at https://dfast.nig.ac.jp/.

2.2.2. *Pan and core genome computation and COG assignment*

For the pan- and core genome analysis, Cluster of Orthologous Group (COG) functional categories were used. Gene clusters in 52 *Bifidobacterium* and 178 *Lactobacillus* was queried against NCBI-CDD using the Reverse Position-Specific BLAST, and its COG category was assigned with a Perl script "cdd2cog" available at https://github.com/aleimba/bac-genomics-scripts/tree/master/cdd2cog. The entire cluster set is called the pan genome.

A commonly used definition of the "core" genome is to select only those genes which are present in almost all genomes. If this traditional threshold for selecting the core

is used, however, the resulting gene number would sharply decrease with the number of genomes compared, leaving ribosomal functions only [77]. To avoid this, I used the trend of the COG categories to determine the degree of gene conservation among complete genomes. Let us define the notion of *n*-core as follows:

n-core … the set of genes that are conserved in *n* percent of the complete genomes.

In *Bifidobacterium*, the 100-core indicates the genes conserved in all 52 genomes, 98-core, genes conserved in 51 genomes, and so on. I created 10 *n*-cores from 100-core to 83-core (genes conserved in 43 genomes, i.e., 83% of the total set). Then for each of the *n*-cores, I obtained the ratio of COG categories (hereafter COG ratio). Each *n*-core showed a different ratio because the number of genes increased as *n* decreases, and their functions in each core became different. To choose an appropriate *n*-core genome, I first created the "consensus COG ordering" based on the majority rule as follows. All COG categories were ordered according to their abundance for each *n*-core and were assigned their ranks. Then, all COG categories were reordered by their average rank (not the ratio) in the 10 *n*-core. I call this COG ordering computed from the average as the consensus COG ordering. Next, each of *n*-core was compared against the consensus COG ordering and the closest core was chosen.

In the case of *Bifidobacterium*, 773 genes that are present in at least 48 genomes (92-core) were selected as the consensus core. For *Lactobacillus*, genes that are present in at least 172 genomes (97-core) were selected as the consensus core. The consensus core of *Bifidobacterium* and *Lactobacillus* contained 773 and 472 gene clusters, respectively.

2.2.3. *Odds ratio and p-value of genus-specific functions*

Statistical analysis was performed to evaluate the relative abundance of protein family among *Bifidobacterium* and *Lactobacillus* species. To examine the over and underrepresented functional categories between both species, I calculated the odds ratio (OR). For calculating OR, a two-by-two contingency table was created. There are four parameters in the table. The parameters are explained as follows: (a) the number of protein families among *Bifidobacterium* present in the respective COG category, (b) the number among *Bifidobacterium* absent in the respective COG category, (c) the number among *Lactobacillus* present in the respective COG category, and (d) the number among *Lactobacillus* absent in the respective COG category. Formula used for calculating OR was ad / bc. The COG categories were defined as overrepresented if OR > 1 and underrepresented if OR < 1. From the OR, 95% confidence intervals (CIs) were calculated as exp[ln(OR)+1.96√(1/a+1/b+1/c+1/d)] for the upper limit, and exp [ln(OR)-1.96√(1/a+1/b+1/c+1/d)] for the lower limit. Let *D* be the difference of natural log (ln) of the upper and the lower limits. The standard error SE was computed as *D* / (2 * 1.96), and the *z score* was ln (OR)/SE. The corresponding *p*-value was obtained with the R function 2 * pnorm(-abs(*z score*)).

2.2.4. *Assignment of carbohydrate-active enzymes*

Carbohydrate-active enzymes were identified based on similarity to the Carbohydrate Active Enzymes (CAZy) database. Online tools CAZYmes Analysis Toolkit (CAT) and dbCAN were used manually [78,79].

2.3. Results and Discussion

2.3.1. *General genomic features of genera Bifidobacterium and Lactobacillus*

The genome sequences of 52 (sub)species of *Bifidobacterium* and 178 (sub)species of *Lactobacillus* were retrieved from the NCBI Assembly database. The genome size of *Bifidobacterium* varied from 1.6 Mb for *B. commune* to 3.3 Mb for *B. biavatii* (Figure 2.1a). The approximate G+C content ranged from 53% to 66% (Figure 2.1b) and the approximate number of coding sequences (CDS) ranged from 1200 to 2500 (Figure 2.1c).

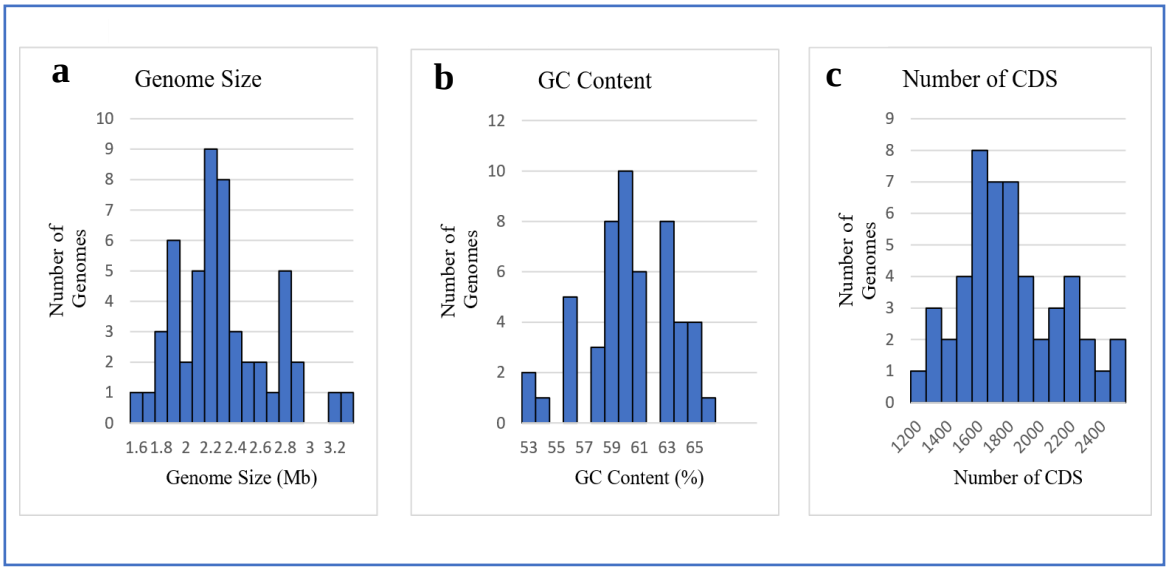


Figure 2.1. Genomic features of genus *Bifidobacterium* [41]

The genome size of *Lactobacillus* ranged from 1.2 Mb for *L. sanfranciscensis* to 3.7 Mb for *L. pentosus* (Figure 2.2a). The G+C content of *Lactobacillus* was lower in comparison with *Bifidobacterium*, ranging from 32% to 57% (Figure 2.2b).

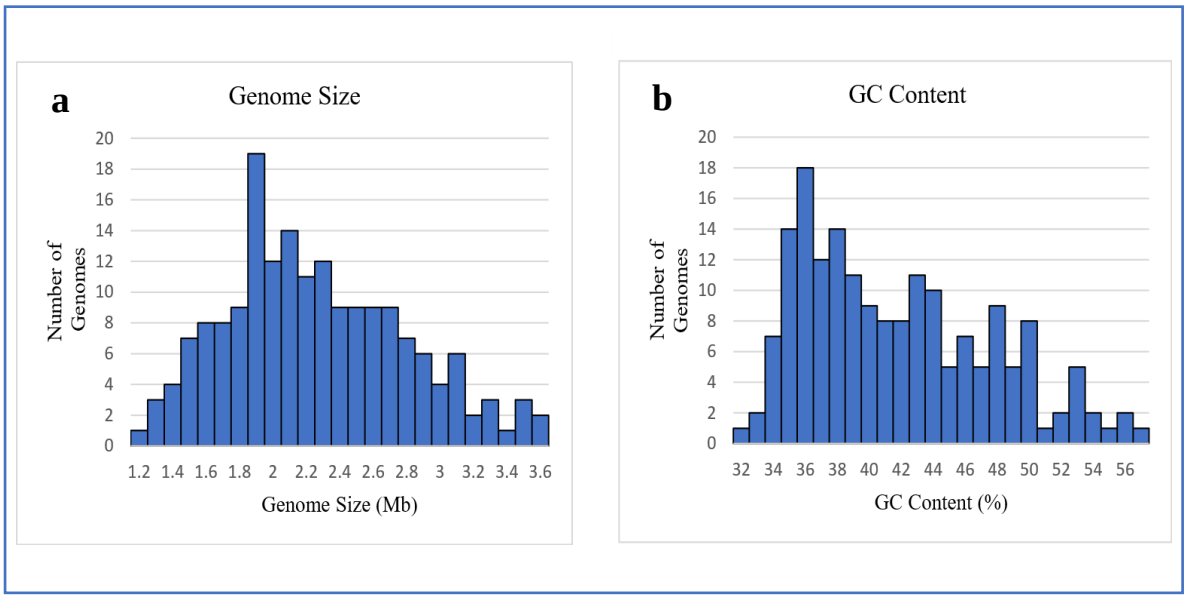


Figure 2.2. Genomic features of genus *Lactobacillus* [41]

Phylogenetically, the two genera are distant. *Lactobacillus* belongs to the phylum Firmicutes as low G+C bacilli, whereas *Bifidobacterium* belongs to the phylum Actinobacteria. However, they share the common niche habitats (animal gut), energy metabolism (lactate

fermentation), and industrial usage as probiotic species. We shall see their metabolic similarities through their genome-scale analyses.

2.3.2. *COG comparison of pan and core genomes of Bifidobacterium and Lactobacillus*

Pan and core genomes for the 52 *Bifidobacterium* species and 178 *Lactobacillus* species were determined as described in Methods section. The pan genome of *Bifidobacterium* and *Lactobacillus* included 16,232 and 45,038 genes clusters, respectively. Functional categories for all the gene clusters for both genera were assigned according to the COG classification. Almost 30% of the pan genome (9,283 gene clusters in *Bifidobacterium* and 26,944 clusters in *Lactobacillus*) were functionally unknown and assigned no COG category (Figure 2.3). This trend is common to many other bacteria; even for *Escherichia coli*, close to 30% of its genes are functionally unknown.

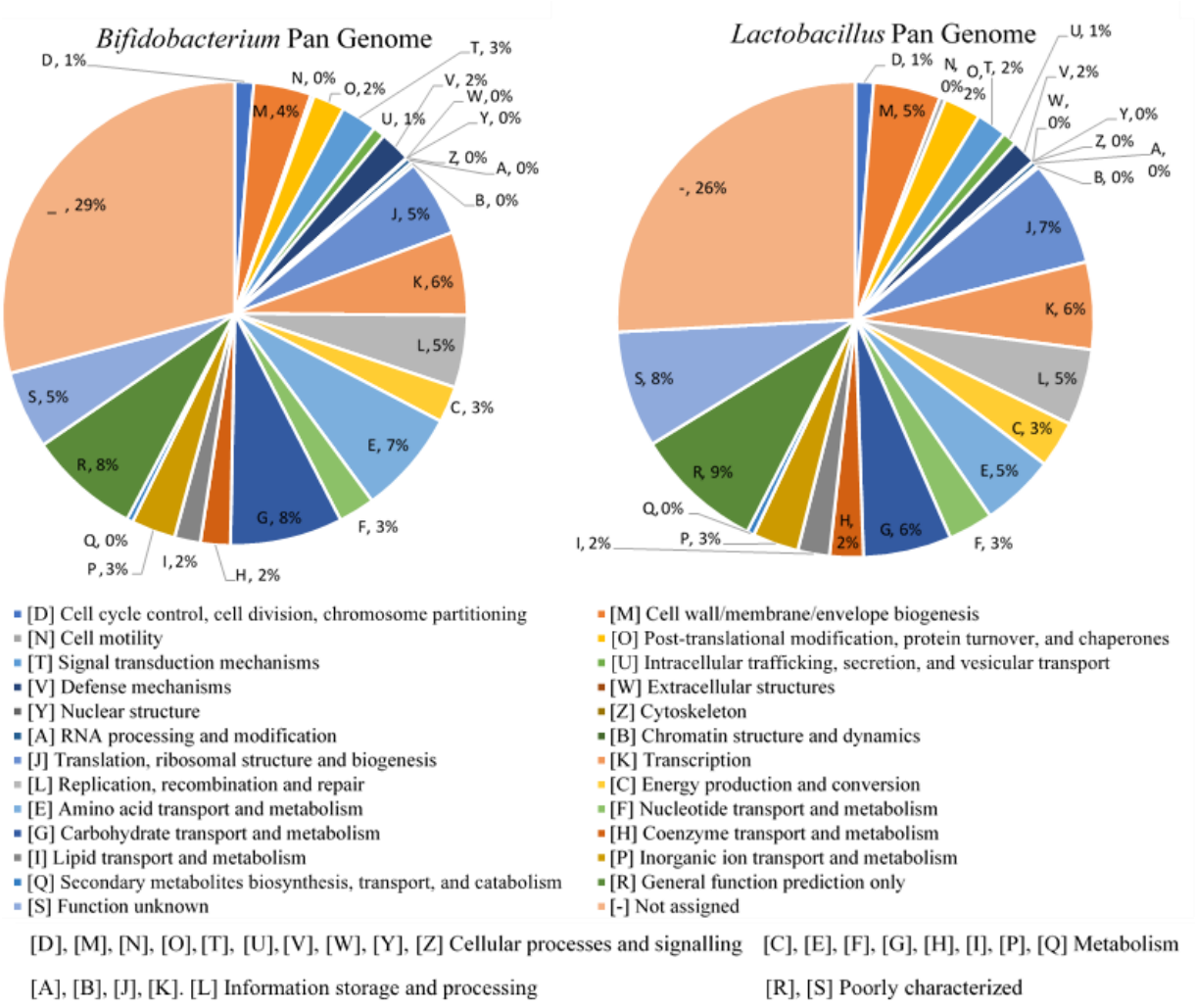


Figure 2.3. COG statistics of the pan genome of *Bifidobacterium* and *Lactobacillus* [41]

The overall COG classification of the pan genome was strikingly similar between the two genera. For both, the highest fractions were identical and the ratio of categories were also identical. Slightly different were the “amino acid transport and metabolism” (E) of 7% vs 5%, and “carbohydrate transport and metabolism” (G) of 8% vs 6% in *Bifidobacterium* and *Lactobacillus*, respectively.

On the other hand, the COG ratios of the core genomes are different. In Figure 2.4, the pie chart of the 92-core (48 out of 52) for *Bifidobacterium* and the 97-core (172 out of 178) for *Lactobacillus* are shown. In the core genome of *Bifidobacterium*, more metabolism related genes, especially “amino acid transport and metabolism” (E), are enriched. In *Lactobacillus* more information storage and processing genes, especially “translation, ribosomal structure and biogenesis” (J), are enriched. In the *Bifidobacterium* core genome, the ratio of “carbohydrate transport and metabolism” (G) is less than in its pan genome. This means that the carbohydrate genes differ from one another within the genus.

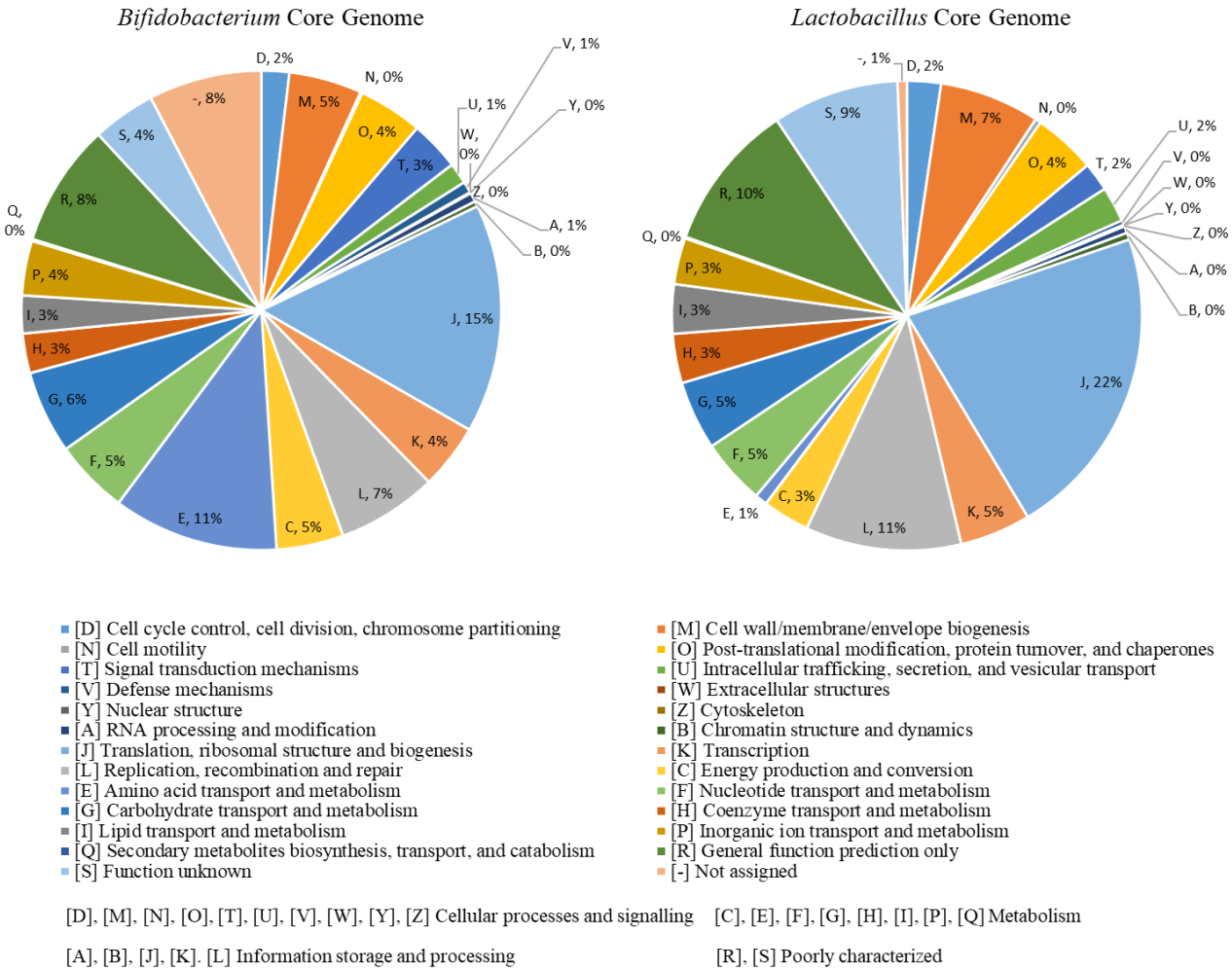


Figure 2.4. COG statistics of the core genome of *Bifidobacterium* and *Lactobacillus* [41]

2.3.3. *Statistical background of the consensus COG ordering*

The difference between pan and core genomes depends on the definition of the core, but there has been no straightforward definition of the core genome. No particular strain such as the type strain necessarily reflect the core, either. The core genome should be an ideal set of genes that represent the respective genus. Intuitively, the 100-core does not reflect the true core because the strict criterion filters too many genes out. On the other hand, the 80-core does not reflect the true core either because the threshold is too relaxed, allowing many auxiliary genes to come in. The ideal core is therefore in-between. To justify our definition of the core genome, let us show the trend of newly added gene functions (COG categories) in each *n*-core (Table 2.1). As *n* decreases from 100, the size of *n*-core increases and genes of different functions are newly included. Each *n*-core is considered a point in the 26-dimensional space (26 is the number of COG functional categories), where the value in each axis is its rank; therefore, no two axis share the same value. Our aim is to find an optimal point in this space from a limited number of sampling ranging from 100-core to 83-core. Under this problem formulation, the easiest way to estimate the optimal is to compute an average rank for each axis independently (note that the averages are not necessarily integers and multiple axes may share the same value), and then to choose the *n*-core nearest to the average (in this way I avoid the problem of indeterminate ranks). The remaining problem is the extent of sampling points. From biological perspective, I considered that 83-core was an appropriate limit to quit sampling because the number of added genes became negligible as shown in Table 2.1.

Table 2.1. The rank (table row) and the number of different COG functional categories in the 10 *n*-cores of *Bifidobacterium*. Color coding: dark … 10 or more gene additions; gray … 10 > and >= 5 gene additions; light gray … 5 > and >=2 gene additions. Some gene numbers are fractions because the same gene may obtain multiple COG categories [41].

100-core (52)	98-core (51)	96-core (50)	94-core (49)	92-core (48)	90-core (47)	88-core (46)	87-core (45)	85-core (44)	83-core (43)
J 76	E 29.5	- 16	E 16.8	- 8	K 5.5	G 9.3	G 5	G 9.3	G 3
L 25.5	J 27	E 14.5	M 6	J 5.5	G 3	I 4	S 5	R 2	V 2
R 24.6	R 19.5	R 10.5	G 6	P 5.5	S 3	G 3	E 4.3	S 2	K 2
O 21.5	- 15	L 9.5	C 5	R 5	R 2.5	P 3	F 3	- 2	M 1

Table 2.1. (Continued)

G 21.3	S 13	J 8	R 4.5	L 4	V 2	R 3	R 3	D 1	O 1
- 21	L 12.8	F 7.5	I 4.3	- 4	E 2	- 3	H 2.5	M 1	E 1
M 20.5	G 11	P 6	F 4	T 3	F 2	M 2	K 2	J 1	F 1
K 19	C 9	K 5	P 3.5	K 3	O 1.5	J 1.5	I 2	C 1	P 1
E 19	M 7.5	S 5	J 3	S 3	L 1.5	E 1.5	- 2	F 1	R 1
F 17.3	F 7.5	H 4.5	S 3	F 2.5	H 1.5	F 1.5	L 1	E 0.5	S 1
C 16.8	O 7	G 3	- 3	M 2	P 1.5	D 1	P 1	H 0.5	D 0
T 12	I 6.3	T 2.5	T 2.5	O 2	D 1	T 1	M 0.5	N 0	N 0
H 10	T 6	C 2.2	D 2	C 2	M 1	V 1	N 0.5	O 0	T 0
P 9	K 5.5	M 2	O 2	G 2	T 1	C 1	U 0.5	T 0	U 0
S 9	U 5	I 2	H 1.8	D 1	J 1	A 0.5	J 0.3	U 0	W 0
I 7	P 4.5	D 1.5	K 1.5	H 1	I 1	L 0.5	D 0	V 0	Y 0
D 6	D 4	O 1	U 1	N 0	N 0.5	N 0	O 0	W 0	Z 0
U 4.5	H 3.3	V 1	A 0.5	U 0	A 0.5	O 0	T 0	Y 0	A 0
A 3.8	V 2	N 0.5	L 0.5	V 0	K 0.5	U 0	V 0	Z 0	B 0
V 3	B 0.5	U 0.5	N 0	W 0	U 0	W 0	W 0	A 0	J 0
B 2.3	N 0	A 0.5	V 0	Y 0	W 0	Y 0	Y 0	B 0	L 0
Q 1	W 0	W 0	W 0	Z 0	Y 0	Z 0	Z 0	K 0	C 0
N 0.5	Y 0	Y 0	Y 0	A 0	Z 0	B 0	A 0	L 0	H 0
Y 0.5	Z 0	Z 0	Z 0	B 0	B 0	H 0	B 0	I 0	I 0
W 0	A 0	B 0	B 0	I 0	C 0	Q 0	C 0	P 0	Q 0
Z 0	Q 0	Q 0	Q 0	Q 0	Q 0	S 0	Q 0	Q 0	- 0

2.3.4. *Relative representation of core genes clusters in Bifidobacterium*

To rigorously identify the over- and underrepresented COG categories in the *Bifidobacterium* core, the odds ratio and *p*-value were computed. Among 777 COG categories analyzed, 359 were overrepresented and 418 were underrepresented in *Bifidobacterium* compared with *Lactobacillus*. Exclusively present and absent COG categories in *Bifidobacterium* were 339 (OR = infinite) and 143 (OR = 0), respectively. COG database has four major functional categories, among which "metabolism" was significantly overrepresented in *Bifidobacterium* with the *p*-value < 0.0001. The other three were underrepresented. Two of them, "information storage and processing" and "poorly characterized," showed the *p*-value of < 0.0001 and 0.0053, respectively. The underrepresentation of the last category "Cellular processes and signaling" was not significant (Figure 2.5a).

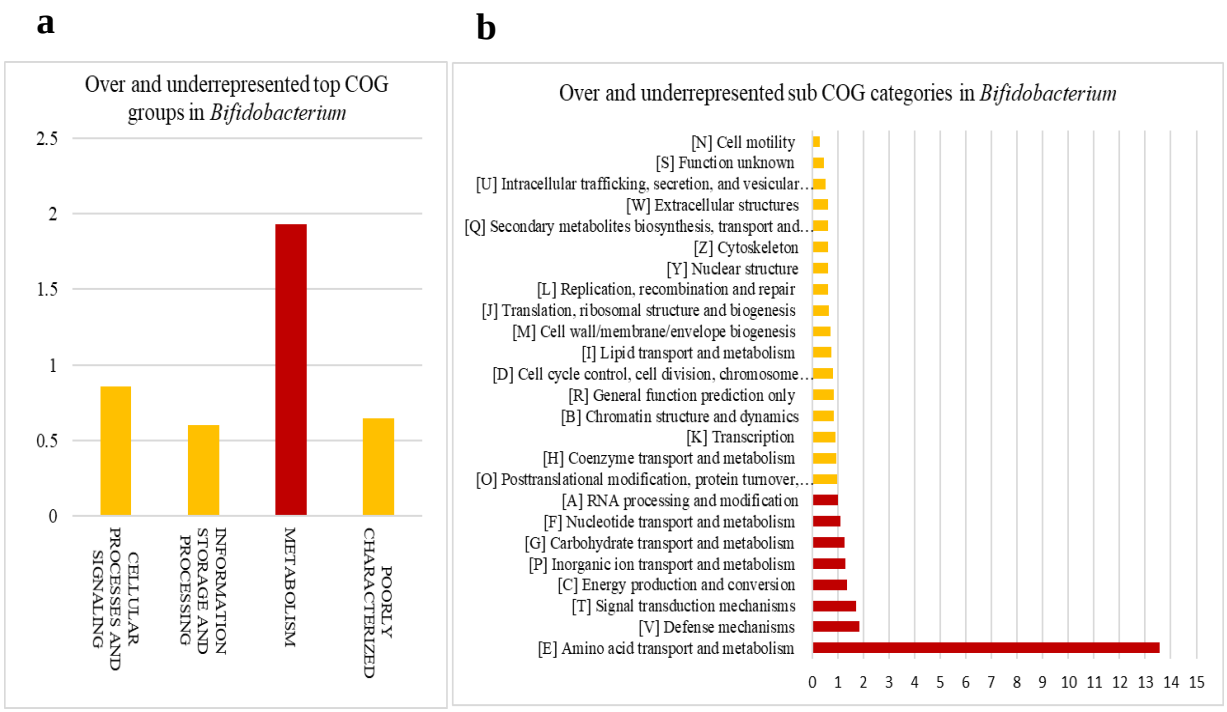


Figure 2.5. Over and Underrepresented COG categories. **(a)** Top four COG groups **(b)** 25 COG subcategories [41].

The four major COG categories were further divided into 25 subcategories. Of the 25 subcategories, 17 were underrepresented and 8 were overrepresented (Figure 2.5b). Among underrepresented categories, J (Translation, ribosomal structure and biogenesis; p-value = 0.0028), L (Replication, recombination and repair; p-value = 0.014) and S (Function unknown; p-value = 0.0017) were statistically significant (Table 2.2). Among overrepresented categories, only E (amino acid transport and metabolism) was significant with the *p*-value < 0.0001. No other category was statistically over- or underrepresented.

Table 2.2. Significantly over- and underrepresented COG categories.
Color coding: dark … significantly overrepresented; gray … significantly underrepresented [41].

Category	a	b	c	d	Odds Ratio	P-value	95% CI
Cellular processes and signalling	145	628	100	372	0.86	0.296	0.64-1.14
Information storage and processing	246	527	205	267	0.61	p<0.0001	0.47-0.77

Table 2.2. (Continued)

Metabolism	310	463	121	351	1.94	p<0.0001	1.51-2.49
Poorly characterized	106	667	93	379	0.65	0.0053	0.47-0.87
[C] Energy production and conversion	40	733	18	454	1.38	0.2707	0.77-2.43
[D] Cell cycle control, cell division, chromosc partitioning	15	758	11	461	0.83	0.641	0.37-1.82
[E] Amino acid transport and metabolism	98	675	5	467	13.56	p<0.0001	5.47-33.56
[F] Nucleotide transport and metabolism	42	731	23	449	1.12	0.6664	0.66-1.89
[G] Carbohydrate transport and metabolism	47	726	23	449	1.26	0.3705	0.75-2.10
[H] Coenzyme transport and metabolism	25	748	16	456	0.95	0.8813	0.50-1.80
[I] Lipid transport and metabolism	25	748	20	452	0.76	0.359	0.41-1.37
[J] Translation, ribosomal structure and biogei	124	649	108	364	0.66	0.0028	0.48-0.85
[K] Transcription	43	730	29	443	0.9	0.67	0.55-1.46
[L] Replication, recombination and repair	62	711	58	414	0.62	0.014	0.42-0.90
[M] Cell wall/membrane/envelope biogenesis	40	733	33	439	0.73	0.1871	1.45-1.16
[N] Cell motility	2	771	4	468	0.3	0.1695	0.05-1.66
[O] Posttranslational modification, protein tur chaperones	35	738	22	450	0.97	0.9131	0.56-1.67
[P] Inorganic ion transport and metabolism	32	741	15	457	1.32	0.389	0.70-2.45
[Q] Secondary metabolites biosynthesis, trans and catabolism	1	772	1	471	0.61	0.727	0.03-9.7
[R] General function prediction only	73	700	52	420	0.84	0.3706	0.57-1.22
[S] Function unknown	33	740	41	431	0.47	0.0017	0.29-0.75
[T] Signal transduction mechanisms	33	740	12	460	1.71	0.1172	0.87-3.34
[U] Intracellular trafficking, secretion, and vesicular transport	13	760	15	457	0.52	0.0893	0.24-1.10
[V] Defense mechanisms	6	767	2	470	1.84	0.457	0.36-9.14
[A] RNA processing and modification	10	763	5	467	1.22	0.7135	0.41-3.60
[B] Chromatin structure and dynamics	7	766	5	467	0.85	0.7878	0.26-2.70
[W] Extracellular structures	0	773	0	472	0.61	0.8054	0.01-30.83
[Y] Nuclear structure	1	772	1	471	0.61	0.727	0.03-9.77
[Z] Cytoskeleton	0	773	0	472	0.61	0.8054	0.01-30.83

Specific COG categories that were overrepresented in *Bifidobacterium* relative to *Lactobacillus* mostly include COGs that were involved in host interaction, stress response and complex compounds metabolism (Table 2.3). In terms of gene multiplicity, glycosidases and galactosidases were multiply retained in many genomes. The table also suggests how *Bifidobacterium* is associated with host and adapted to the habitat of gastrointestinal tract.

Many categories were within carbohydrate transport or metabolism (category G) to compete and survive in the intestine. The presence of plasminogen binding, mucin binding and proteins involved in formation of capsular or exo-polysaccharides (EPSs) has been well reported as the characteristics of *Bifidobacterium* [80]. Other overrepresented COGs included a number of protein families to adapt to the adverse intestinal environment such as the bile and stress resistance. One important protein found in the core was LuxS, which is involved in biofilm formation and help *Bifidobacterium* in gut colonization and pathogen protection [81]. Immunoreactive proteins like aspartokinase and surface antigens which contact with host cells and modulate immune response were also identified. These proteins as well as ABC-type sugar transporters and transaldolases have been experimentally documented as important in this genus [82]. Without differential analysis, typical representative cores would be unknown functions (category -, S or R) or ribosomal functions (category J) (Figure 2.4). Unbiased selection of the above functions validates the appropriateness of our core definition and the importance of comparative analysis.

Table 2.3. Overrepresented COG categories in *Bifidobacterium.* Redundancy refers to the total number with possible multiplicity in the same genome. Gene functions are based on the annotation in our reference library [41].

COG Category	Odds Ratio	Number of genomes	Redundancy	Function
[E] COG1113 Gamma-aminobutyrate permease and related permeases	infinite	48	48	Interaction of gut microbiota with the macroorganism
[E] COG0527 Aspartokinases	infinite	49	49	Immunoreactive proteins
[O] COG0265 Trypsin-like serine proteases	infinite	51	51	Involved in stress response eg bile response
[G] COG3345 Alpha-galactosidase	infinite	49	67	Involved in complex carbohydrate metabolism (glycosyl hydrolases)
[M] COG1247 Sortase and related acyltransferases	infinite	49	49	Binding with host
[T] COG1854 LuxS protein involved in autoinducer AI2 synthesis	infinite	51	51	Biofilm formation and host colonization

Table 2.3. (Continued)

[R] COG3942 Surface antigen	infinite	52	52	Immunomodulatory activity
[G] COG0021 Transketolase	infinite	52	53	Extracellular proteome
[G] COG0176 Transaldolase	infinite	52	52	Mucin binding capability and aggregation factor
[G] COG0366 Glycosidases	infinite	51	72	Involved in complex carbohydrate metabolism (glycosyl hydrolases)
[G] COG0166 Glucose-6-phosphate isomerase	infinite	52	53	enzyme involve in bifidus pathway (glucose metabolism)
[E] COG0334 Glutamate dehydrogenase/leucine dehydrogenase	infinite	49	49	Plays central roles in nitrogen metabolism
[M] COG1215 Glycosyltransferases	infinite	52	52	Involved in EPS production
[G] COG0033 Phosphoglucomutase	infinite	52	52	Galactose metabolizing enzyme
[G] COG0580 Glycerol uptake facilitator and related permeases (Major Intrinsic Protein Family)	infinite	52	52	Sugar transport (Non-PTS sugar transport system)
[GM] COG1134 ABC-type polysaccharide/polyol phosphate transport system	infinite	49	49	Sugar transport (Non-PTS sugar transport system)
[GM] COG1682 ABC-type polysaccharide/polyol phosphate export systems	infinite	49	49	Sugar transport (Non-PTS sugar transport system)
[G] COG3839 ABC-type sugar transport systems	infinite	52	52	Sugar transport (Non-PTS sugar transport system)

It is arguable that the core genes can be computed for any set of strains, e.g., to identify host-specific gene pools. Indeed, Sun *et al*. identified cytochrome d oxidases as the core of bee-specific *Bifidobacterium* and compared the list with other host-specific genes [77]. However, for cross-species comparison as in this study, statistical criterion of the core was preferred and the problem of host-unspecific species (all-rounders) was hard to resolve. The relationship between the core genes and host remains the future problem of our study.

2.3.5. *Comparison of carbohydrate metabolism in Bifidobacterium and Lactobacillus*

One interesting application of the core in the two probiotic species is the characteristics of carbohydrate metabolism [83]. I manually investigated the carbohydrate-active enzymes encoded in the two genera and confirmed the presence of glycosyl hydrolases (GHs), glycosyl transferases (GT), carbohydrate esterases (CE), polysaccharide lyases (PLs) and carbohydrate binding modules (CBM).

According to Carbohydrate Active Enzymes (CAZy) system classification, the pan genome of *Bifidobacterium* included 69 GHs, 12 GT, 6 CE, 26 CBM and no PLs, while that of *Lactobacillus* included 48 GHs, 10 GT, 6 CE, 10 CBM and 15 PLs. Since the number of analyzed genomes in *Bifidobacterium* is less than a third compared to *Lactobacillus*, the number of carbohydrate-active enzymes in the former was larger except for PLs. Most common enzymes in *Bifidobacterium* were the four types of glycosyl hydrolases: GH43 for xylanase and arabinanase, GH25 for muramidase, GH3 for beta-glucosidase, and GH13 for amylase. On the other hand, in *Lactobacillus*, most common were GH73 for peptidoglycan hydrolase, GH25 for muramidase and GH32 for fructan hydrolase. The characteristic types of glycosyl hydrolases reflected the different types of oligosaccharides the bacteria can metabolize: *Bifidobacterium* digests relatively animal oriented sugars and *Lactobacillus,* plant oriented. No PLs in *Bifidobacterium* is noteworthy, but we are skeptical because multiple evidences showed that they can metabolize uronic acid containing polysaccharides such as pectins and hemicelluloses [84]. These uronic acid containing polysaccharides come from bacteria, plant, or animal, and it is unlikely that the bacteria do not metabolize them.

2.4. Conclusion

In this study I created a free, curated reference library for genus *Bifidobacterium*, which enable any user the accurate and consistent annotations for newly sequenced *Bifidobacterium*. In the orthologous gene cluster analysis, the pan genomes of *Bifidobacterium* and *Lactobacillus* consisted of 16,232 and 45,038 clusters, respectively. From among them, core genes in each genus were selected based on a statistical definition of core genome: for *Bifidobacterium* gene clusters present in at least 92% of genomes and for *Lactobacillus*, 97%. Through its comparative analysis with another probiotic genus *Lactobacillus*, their metabolic characteristics were revealed: protein families overrepresented in *Bifidobacterium* were mostly

involved in complex sugar metabolism host interaction, and stress responses. These functions were in good agreement with known literature data. The analysis also showed more niche adjusted metabolic activities such as broad adaptability for amino acids and polysaccharide metabolism in *Bifidobacterium*. The relative absence of polysaccharide lyases was shown but further analysis is required to conclude the metabolic ability on polysaccharides or host-specificity.

CHAPTER 3

Comparative Analysis of *Bifidobacterium. bifidum* TMC3115 strain and Insight into its Immunomodulatory Role

3.1. Introduction

Among the bifidobacteria, *B. bifidum* species are of special interest because of their abundance in the gut of infants, specially breast-fed infants [85]. The ability of *B. bifidum* species to degrade mucin and the presence of pili, specially the sortase-dependent pili helps them to interact with host and adapt to the gastrointestinal habitat [86]. These species were found to be involved in maturation of immune system which is not fully developed at the time of the birth. *B. bifidum* compared to other *Bifidobacterium* species have shown high production of IL-17 cytokine [87]. *B. bifidum* species have their role in immunomodulatory activity and in strengthening of innate immune system during the host colonization. *B. bifidum* PRL2010 shows high production of interleukin 6 (IL-6) and IL-8 cytokines, probably by NF-κB activation and suggested to modulate the immune response of the host [88]. One of the important benefits of probiotic bacteria is modulation of host immune system. Probiotic bacteria can prevent various severe diseases like ulcerative colitis, allergy and atopic diseases by stimulating the immune system [89,90]. They impart beneficial effects by regulating the production of anti and proinflammatory cytokines and balance of T helper (TH1)/TH2 [91].

The exact mechanism behind the impact of probiotics on host immune system is yet not clear. Possibly the extracellular factors play vital role in bacterial host interaction. For the survival, colonization, immune stimulation and probiotics characteristics of commensal bacteria, different structures like extracellular proteins, pili and teichoic acids plays important role. Extracellular proteins, which are either secreted or present on the surface of the cell maintain the homeostasis in the gastrointestinal tract (GIT) by different mechanism like adhesion to mucin and epithelial cells, modulating immune cells and cross talking with host cells. The extracellular proteins present in probiotic bacteria like *Lactobacillus* were found to be involved in interaction with host cells [92]. In altering immune system, colonization and

adhesion with host, pili structure specially sortase dependent pili are important. In *Lactobacillus rhamnosus* GG and *B. bifidum* PRL2010 sortase dependent pili were considered to play vital role in adherence and immunomodulation in host [93,51].

The significant potential benefits of *Bifidobacterium* to its host are strain-dependent. Comparative analysis among various strains can reveal the important features for each strain. Comparative genomics among the bifidobacterial complete genomes has revealed a high degree of synteny among the entire genomes of a taxa, however for some taxa inversions, DNA insertions and deletions are also commonly detected [94]. Comparative analysis can reveal the core and unique genomic regions providing insightful genomic information.

In the present study, the *B. bifidum* strain, TMC3115 was analysed. *B. bifidum* TMC3115 strain was isolated from healthy infant. It can adhere to intestinal epithelial cells and mucosa without the inflammation [95]. The strain shows high inhibitory effects on IgE-mediated allergic inflammation [96]. Studies show that by affecting the function of intestinal epithelium and immunity TMC3115 strain can modulate intestinal microbiota and it could also protect the host animals from antibiotic side effects [97].

The objective of this study was to explore the genomic features of this important strain which are not well characterized at present. The genomic structure, unique genomic features and probiotic characteristics specifically the immunomodulatory role of this strain was examined. A comparative genomic approach was used for this purpose. The detailed analysis focusing on genome synteny and genomic features (extracellular proteins and pili like structures) having role in host-microbe interaction and immunomodulation was done.

3.2. Methods

3.2.1. *Annotations and COG assignment*

In the analysis 10 complete genomes including the TMC3115 genome and 22 draft genomes of *B. bifidum* were used. The reference library for *Bifidobacterium* in DFAST was used for annotating all the genomes [76]. Cluster of Orthologous Group (COG) functional categories were assigned by querying all the proteins against NCBI-CDD using the Reverse Position-Specific BLAST, and COG categories were assigned with a Perl script “cdd2cog” available at https://github.com/aleimba/bac-genomics-scripts/tree/master/cdd2cog.

3.2.2. *Genome synteny*

To examine the genome synteny, the whole genome alignment of TMC3115 strain with other complete genomes was done. The dot plot alignment was done using GEPARD [98]. To visualize the potential inversion detected by dot plot results, the complete genomes were visualized using a circular genome visualizer [99]. Further the inversion breakpoints were identified using the algorithm proposed for rearrangement identification in multiple genomes by Noureen et al., [100]. The repeat sequences around the inversion breakpoints were identified using Unipro Ugene software version 35.1[101]. To find the effect of inversion on replication site, oriC and terC sites were determined. The oriC site was identified using Ori-Finder 2 [102] and terC site was find based on GC skew analysis using GenSkew (https://genskew.csb.univie.ac.at/) and based on the consensus sequence for actinobacterial genomes [103]. The flanking regions of inversions are checked for the presence in the genomic islands (GI) using IslandViewer4 webserver [104].

3.2.3. *Extracellular proteins identification*

For identification of extracellular proteins, first classically secreted proteins were screened out by checking the presence of signal peptides using SignalP3 [105]. Proteins which were not detected as classically secreted were than checked by SecretomeP [106]. All the proteins detected as secreted proteins from both methods were than further checked for presence of transmembrane helices with TMHMM2 [107]. Proteins detected to have three or more transmembrane helices were excluded and remaining proteins were detected as potential extracellular proteins. The detailed strategy used for identifying the extracellular proteins is shown in Figure 3.1. Further the cellular and subcellular localization of the identified proteins were predicted by Psortb version 3 [108] and LocateP [109]. LipoP [110] was used to predict lipid anchor motifs. Functional classes were assigned on the basis of COG functional classes and domains were identified by Pfam [111].

3.2.4. *Identification of sortase dependent pili*

Pili-encoding proteins were identified based on amino acid similarity by performing BLAST analysis. Further detailed in silico analysis of motifs and domains present in pilin subunits was done. For this Sec-dependent secretion signal, sortase recognition site (CWSS motif), the pilin-like motif (TVxxK) and E box were checked [112].

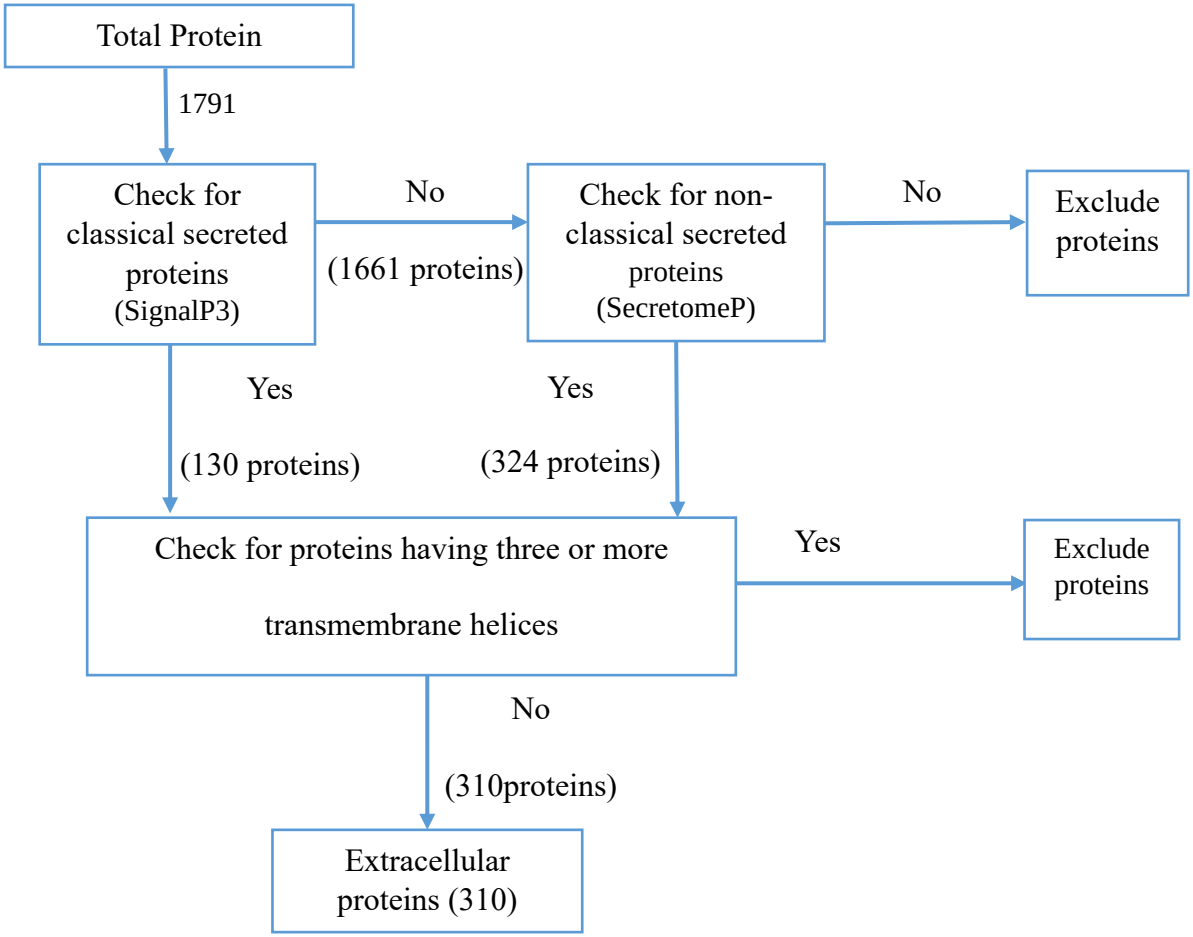


Figure 3.1. Schematic representation of the scheme followed for identification of extracellular proteins

3.3. Results and discussion

3.3.1. *Comparative analysis of B. bifidum TMC3115 strain*

3.3.1.1. *Genomic features*

B. bifidum TMC3115 have a genomic size of 2178894 bp with the GC content of 62.8 %. A total of 1791 coding DNA sequences were predicted with an average length of 346 bases constituting 85.3 % of the genome. The genome consists of 191 pseudogenes and 53 tRNA genes. GC-skew analysis identified that the oriC site was located proximal to dnaA gene at ~1.6 mb and terC at ~ 0.32 mb. The general characteristics of TMC3115 strain along with other nine complete genomes of *B. bifidum* are shown in Table 3.1.

Table 3.1. Genome features of *B. bifidum* TMC3115 strain and other 9 *B. bifidum* complete genome

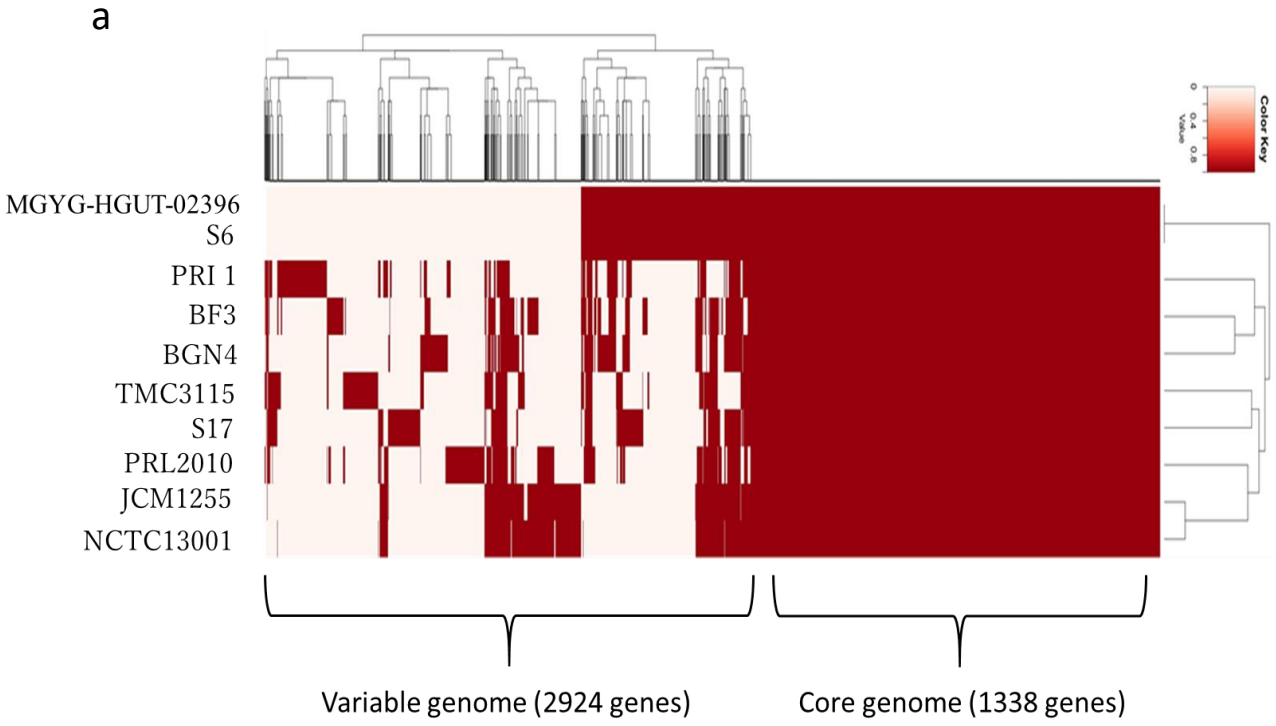
Genbank Accession	Strain names	Origin	Genome Size	GC Content	Number of CDS	Number of tRNAs	Number of rRNAs	Number of unique genes	Number of insertion sequences
AP018132.1	TMC3115	Infant	2178894	62.8	1791	53	0	105	45
CP010412	BF3	Infant	2210370	62.7	1816	53	0	40	25
NZ_CP018757	PRI 1	adult	2243572	62.6	1857	53	0	147	46
NZ_CP022723	S6	adult	2311342	62.7	1915	55	0	0	34
NZ_LR134344	NCTC13001	Infant	2211032	62.7	1865	54	0	10	30
NZ_LR698991	MGYG-HGUT-02396	adult	2311342	62.7	1915	55	0	0	34
AP012323	JCM 1255	Infant	2211039	62.7	1861	54	0	4	30
CP001361	BGN4	Infant	2223664	62.6	1826	53	0	49	24
CP001840	PRL2010	Infant	2214656	62.7	1864	53	0	110	31
CP002220	S17	Infant	2186882	62.8	1821	54	0	93	22

Functional classification according to cluster of orthologous groups of proteins (COGs) classified 1381 (77.06%) proteins into 26 COG categories. Remaining 410 proteins (22.93%) were not assigned any of the COG category. Most of the proteins are present in the five categories: 150 proteins in category R (General function prediction only), 133 proteins in category J (Translation, ribosomal structure and biogenesis), 126 proteins in category E (Amino acid transport and metabolism), 121 proteins in category L (Replication, recombination and repair) and 117 proteins in category G (Carbohydrate transport and metabolism). Among the top COG categories much more of the metabolism related proteins were present. The

distribution of proteins in COG top categories and sub categories is shown in Supplementary Figure 3.1.

Comparative analysis *B. bifidum* TMC3115 was done with other 9 strains having complete genomes. The orthologous clustering of the complete genomes resulted in 1338 core genes and 2924 genes as variable. The distribution of COG categories showed that there is not much difference in the COG classes among the strains. Further the comparative analysis revealed that some of the strains don't have any of the unique gene while some strains have more than 100 unique genes. The genome of TMC3115 have 105 of these unique genes which mostly include the hypothetical genes and the transposases (Supplementary Figure 3.2).

The whole genome comparison taking TMC315 strain as a reference reveal the regions in the genome which are variable (present in some strains) and unique (present only in TMC3115). In total 10 variable regions and 2 unique regions were identified. The genes in the variable regions mostly include the transposases and hypothetical proteins (Figure 3.2). Other classes of genes in these variable regions include integrase, restriction modification system, teichoic acid synthesis, glycosyl transferases, magnesium transporters, beta-galactosidase and metallo-beta-lactamase. The presence of the genes related to transposases, integrase and restriction modification system among the variable regions suggest that horizontal gene transfer (HGT) is potentially one of the driving forces of evolution causing the diversification among these *B. bifidum* complete genomes.



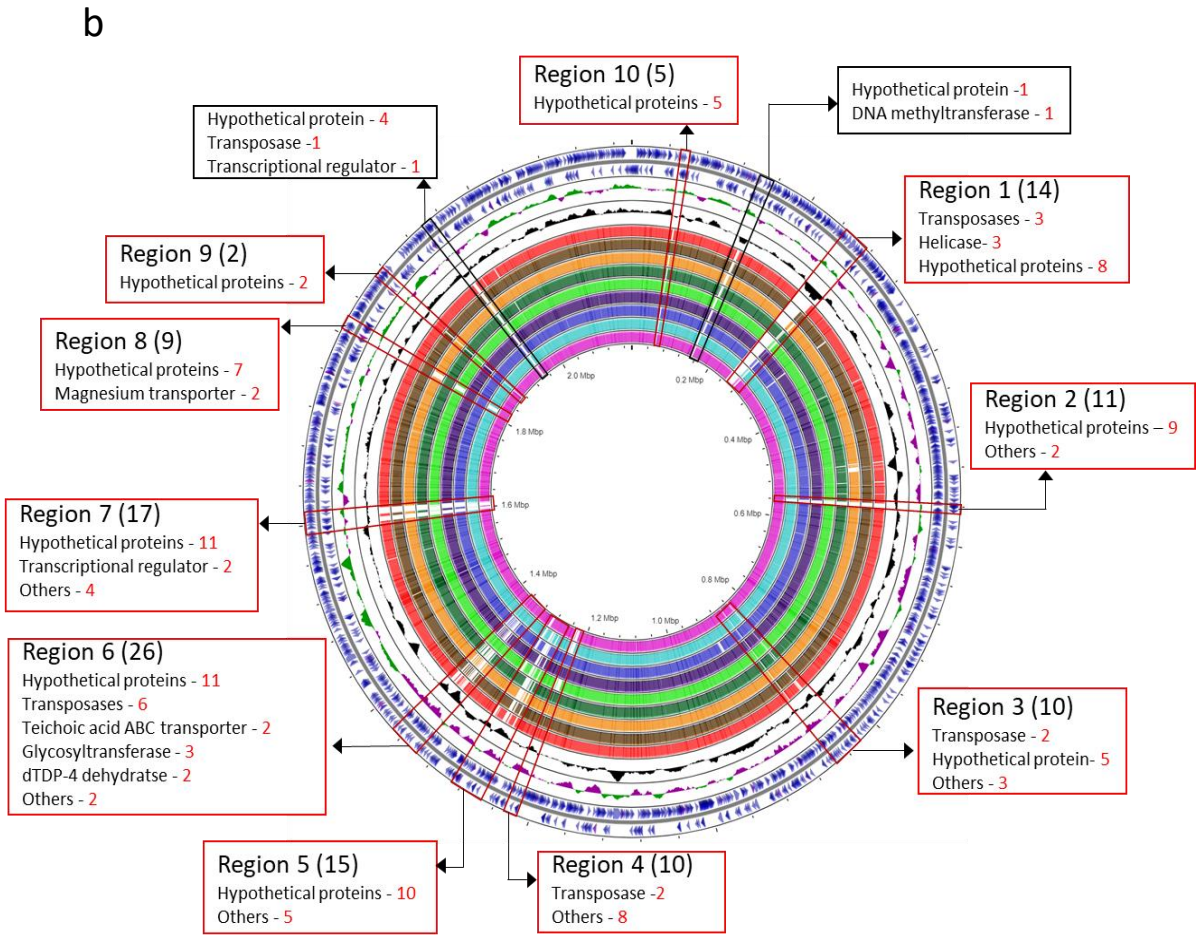


Figure 3.2. Comparative genomics of *B. bifidum* TMC3115. **(a)** The heatmap showing the hierarchical clustering of core and variable genome based on the presence and absence of genes. **(b)** Variable regions in *B. bifidum* complete genomes. Blast-based genome atlas showing the genomic variability among *B. bifidum* complete genomes taking TMC3115 as a reference genome. The regions showing variability are highlighted in red while the regions unique for TMC3115 strain are highlighted in black. The gene information name and their number are shown in red boxes. The single genes are represented as others in the labels. This comparative analysis was done using CGView Server.

3.3.1.2. *Mobilome of TMC3115 strain*

Based on prophage identification by PHASTER, none of the intact prophage elements were identified in TMC3115, however 2 incomplete and one questionable prophage regions were identified. A total of 45 insertion sequences (IS), including 7 IS families: IS3, IS21, IS30, IS91, IS256, IS607 and ISL3 were found in TMC3115 genome. The IS are present in the genome

clustered together forming 7 IS regions. IS256 and IS3 were found to be most frequent insertion family in TMC3115. Figure 3.3 shows the detailed IS regions.

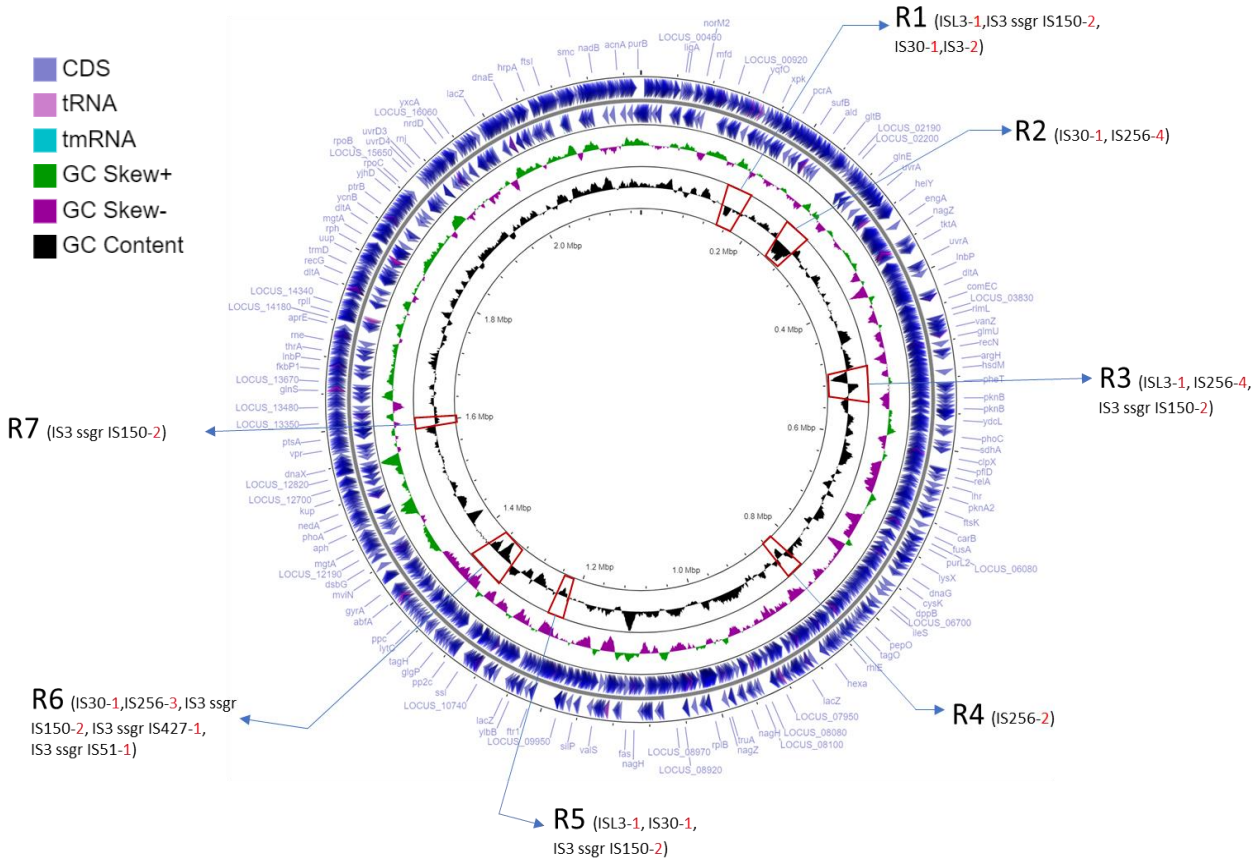


Figure 3.3. The genome map of *B. bifidum* TMC3115. The map shows various features of TMC3115 strain: CDS, tRNA, tmRNA, GC content and GC skew. The regions with the insertion sequences (IS) are highlighted in red. The label shows the type of the IS in each region and the number of IS.

3.3.1.3. *Synteny and genomic architecture*

The whole genome alignment of TMC3115 with other *B. bifidum* complete genomes revealed that TMC3115 strain does not show synteny with other *B. bifidum* genomes (Figure 3.4a and 3.4b). The comparison based on multiple genome visualizer and genome rearrangement analysis revealed a large inversion of ~ 382 kb in TMC3115 strain. The inversion occurs between 1231692 to 1614181bp (~ 382 kb). The detail analysis of the terminal part of the inversion (breakpoints) was done to identify the possible cause of the inversion. Studies shows that repeat sequences and IS can be the possible cause of inversions [113,114].

The analysis of the flanking regions of the inversion reveals the presence of following genomic elements (i) an inverted repeat of 1144 bp, (ii) genomic island having IS elements of family IS3 ssgr IS150, (iii) type II restriction modification system. The detail analysis shows that the possible mechanism behind this identified inversion is DNA duplication. The genomic region encoding for transposase and integrase genes is duplicated in an inverted orientation at the other end of the inversion breakpoint causing the recombination between the two loci. The schematic view of the inversion caused by duplication is shown in Figure 3.4c.

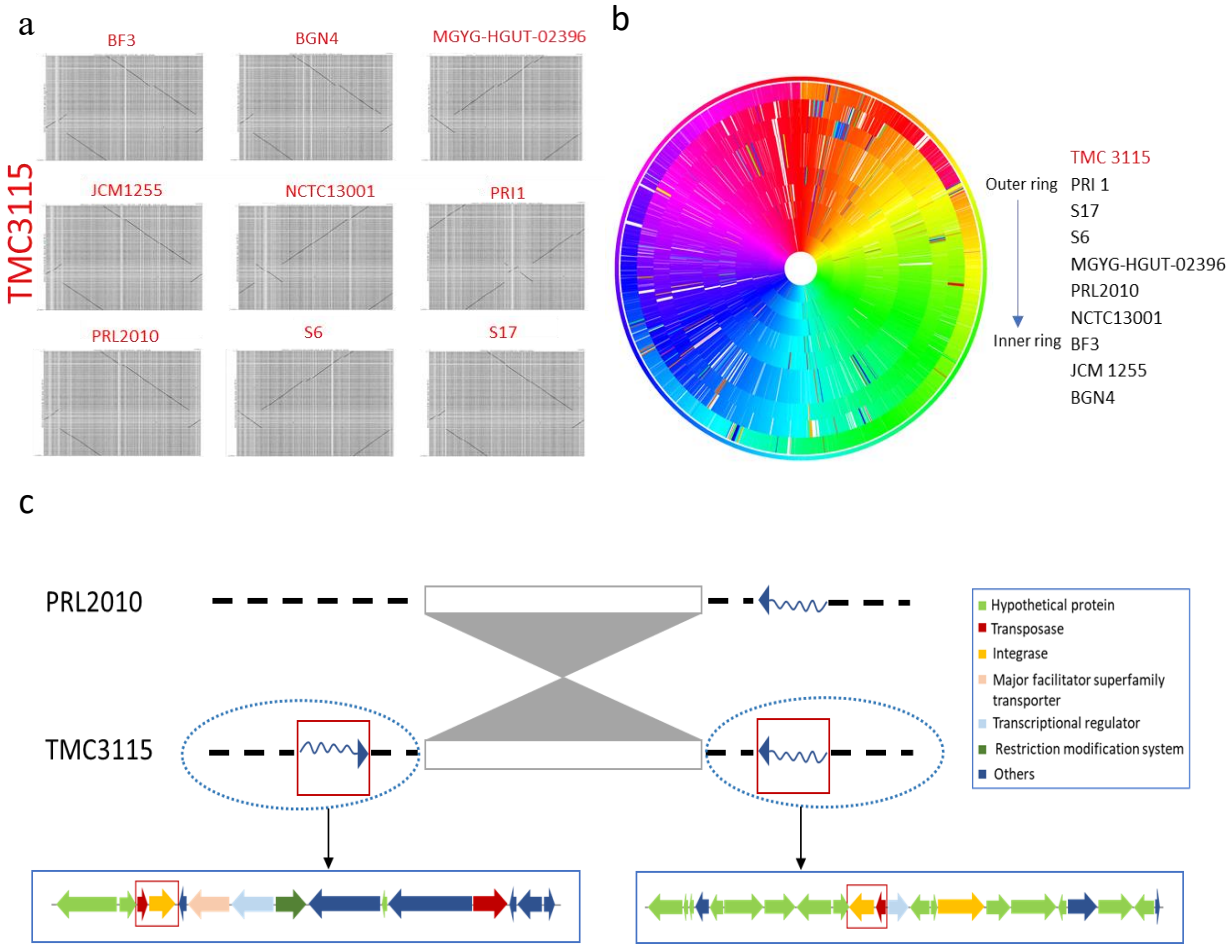


Figure 3.4. Genome synteny and architecture of *B. bifidum* TMC3115. **(a)** Dotplot alignment of TMC3115 with all the complete *B. bifidum* genomes. The dotplot shows that TMC3115 strain does not show synteny with other strains of *B. bifidum*. **(b)** Circular view based on orthologous clusters. In this view the genes are color coded by genomic position of the cluster they belong to. The outermost ring representing TMC3115 strain shows the shift in the colors which correspond to the presence of genome rearrangement in TMC3115 strain. **(c)** The schematic view of inversion caused by duplication. The genomic region represented by blue

wavy arrow in PRL2010 strain is duplicated in inverted orientation in TMC3115 strain causing the genome rearrangement in TMC3115. The gene map for the genes present around the breakpoint regions is shown in blue boxes. Among which the genes encoding for transposase and integrase highlighted in red box are duplicated at both breakpoints.

The inversion is present near the oriC region. It changed the symmetry between the oriC and terC by shifting the terC region to 203 kb before the symmetry center (Figure 3.5). Often the large shift from the replication symmetry can have some effects on the strain growth or genome integrity however studies show that sometimes even a big shift of the termination has no serious effects. They might have some positive impacts in certain environmental conditions [115,116].

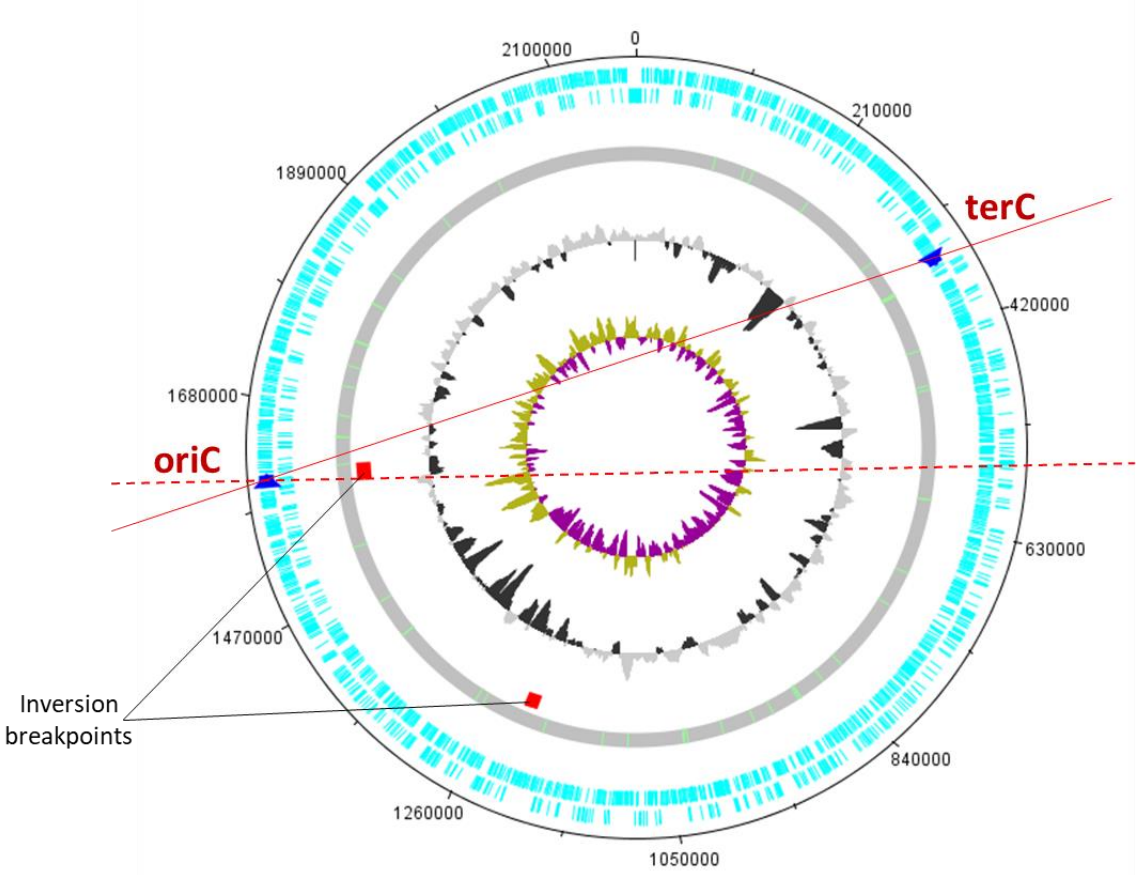


Figure 3.5. The circular genome map of TMC3115 showing the oriC and terC sites. From the inner most circle: Circle 1 shows the GC skew. Values >0 are in gold and <0 are in purple. Circle 2 illustrate percentage GC plot. Circle 3 indicates tRNA and rRNA highlighted in green. Circle 4 and 5 shows the forward CDS and reverse CDS. The inversion breakpoint and its flanking region is highlighted in red. The red dotted line shows the replication symmetry while the plain red line shows the shift from the replication symmetry.

Among the other complete genomes, the strains of PRI 1 show also inversions of ~ 171 kb and the inversion occurs between 1314345 to 1143220 bp. The detail analysis of region around inversion shows that the same phenomenon of duplication of genes in inverted orientation as observed in TMC3115 has caused this inversion also. The genomic region encoding for IstB-like ATP-binding protein and transposase belonging to family of IS21 is duplicated in an inverted orientation.

3.3.2. *Host interaction and immunomodulatory role of TMC3115 strain*

To interact with the host, microorganisms exhibit specific strategies. Although bifidobacteria impart beneficial effects on host health yet the mechanism that how they attach to their host intestinal epithelial and elicit the immune response in still unknown. Certain extracellular structures, secreted enzymes and cell wall components like teichoic and lipoteichoic acid plays an important role in host interaction thereby modulating the immune system [47, 117]. In the genome of TMC3115 we examined these extracellular structures to get insight into its host adaptation and immunomodulatory role.

3.3.2.1. Extracellular proteins

The extracellular proteins can facilitate probiotic characteristics by mediating the interactions with mucosal cells, such as epithelial and immune cells [118] Extracellular protein in bifidobacteria have important role in host interaction and adaptation. Further, in *Bifidobacterium* the extracellular proteins are directly involved in beneficial mechanisms for the host thus important to study the host interactions [119]. Mostly the proteins are secreted to the extracellular space by N-terminal signal peptide known as classical secreted proteins however sometimes the extracellular proteins lack this signal peptide and are secreted by non-classical secretory pathway [120].

In the genome of *B. bifidum* TMC3115, 310 potential extracellular proteins were identified among which 97 proteins were classically secreted and 213 were non-classically secreted proteins. On the basis of functional prediction done by COG classes and Pfam domains identified proteins were categorized into five categories of enzymes, transporters, regulators or signal transduction, unknown and others. Identified extracellular proteins were also characterized on basis of anchor types. Eight classes were found, secreted (21), LPxTG cell-wall anchored (32), lipid anchored (14), N-terminally anchored (30), C-terminally anchored

(1), LysM domain proteins (1), Unkown (9) and Others (202). Table 3.2 shows the anchor types and functional categorization of the potential extracellular proteins of TMC3115 strain.

Table 3.2. Functional categories and anchor types of predicted extracellular proteins

Functional Category	Secreted	LPxTG Cell-wall anchored	Lipid anchored	N-terminally anchored	C-terminally anchored	LysM domain proteins		Unknown	Others	Total Proteins
Enzyme	4	19	1	8	0	0	2		43	77
Transporter	1	0	10	1	0	0	1		12	25
Regulator/ Signal transduction	0	0	0	3	0	0	0		9	12
Unknown	14	7	2	17	0	0	5		96	141
Other	2	6	1	1	1	1	1		42	55
Total Proteins	21	32	14	30	1	1	9		202	310

Among the identified proteins, twenty-one are secreted proteins with most of them having unknown function. Some of the predicted secreted proteins contain binding domains such as F5/8 type C domain, CHAP domain and G5 domain. CHAP domain is previously characterized as important domain playing role in the interaction of bifidobacteria with the host immune system [121]. G5 domain is involved in N-acetylglucosamine binding and has role in biofilm formation in bacteria [122]. Thirty-two of these proteins have LPxTG cell wall-sorting motif for covalently binding to peptidoglycan by sortase [123], most of them are functionally categorized as enzymes. LPxTG cell-wall anchored proteins plays important role in adhesion, host colonization and immunomodulation. They contain 5 pilin proteins for sortase dependent pili including both major (FimA or FimP) and minor (FimB or FimQ) pilin proteins. Proteins like sialidase involved in mucosal surface adhesion [124], fucosidases degrading host-derived glycans [125], chitin protein having role in adhesion were present [126]. Fourteen proteins are lipid anchored in which covalent binding of N-terminal cysteine residue to lipids help them to anchor [127], among these many transporter proteins are present. Among non-covalently

attached proteins 30 N-terminally anchored, 1 C-terminally anchored and LysM domain containing proteins were present.

The identified extracellular proteins contain 29 proteins having important binding domains and play role in host interaction. Further Blast analysis with immunoreactive proteins which showed homology with identified immunoreactive proteins in *B. longum subsp. longum* CCM 7952 and *B. longum subsp. longum* CCDM 372 [128], showed 4 proteins having similarity with these proteins (Table 3.3).

Table 3.3. Homology with the immunoreactive proteins

Protein name and accession	Homologous protein name and Acession	Identity %
BBA56216.1 molecular chaperone DnaK	NP_695712.1 Molecular chaperone DnaK	96
BBA56628.1 cell division ATP-binding protein FtsE	NP_695858.1 Sugar ABC transporter ATP-binding protein	37
BBA56686.1 peptidoglycan synthetase FtsI, penicillin-binding	EDT88982.1 Penicillin-binding protein 3 peptidoglycan synthetase	58
BBA56449.1 30S ribosomal protein S16	30S ribosomal protein S16 Q8G7G1.1	88

3.3.2.2. Sortase dependent pili clusters

Sortase dependent pili clusters were identified on the basis of similarity search and detail analysis of domains and motifs. Three sortase dependent pili clusters were identified in TMC3115 strain, shown in Figure 3.6.

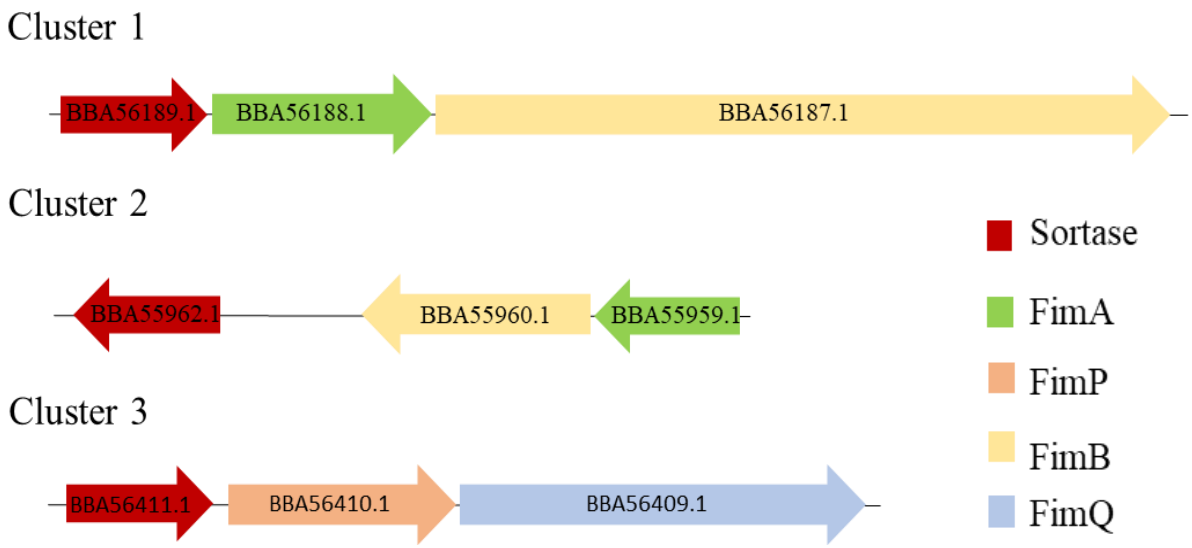


Figure 3.6. Genetic map representing the pili clusters in TMC3115, the arrows show different genes and numbers on arrows represent the locus tags.

The pili clusters have genes encoding major pilin subunit (FimA or FimP) and minor pilin subunit (FimB or FimQ) and sortase gene. Pili are considered important for bacteria host interaction. They have function in adhesion, biofilm formation and immunomodulation [93]. SD pili identified in *B. bifidum* PRL2010 shows immunomodulatory activity [51]. The homology of pili clusters proteins of TMC3115 strain with PRL2010 and the pilin motif, sortase recognition site (CWSS motif) and E box sequence detected for six pilin proteins are shown in Table 3.4. The high homology in proteins of pili cluster of TMC3115 and PRL2010 shows that the proteins in pili cluster in both strains are quite similar so possibly they are also involved in immunomodulatory interaction with host gut cells.

Table 3.4. Homology of pili clusters proteins and the pilin motif, CWSS motif and E box sequence

	Protein name and accession	**Identity**	**CWSS**	**Pilin Motif**	**E box**
Cluster 1	BBA56189.1 sortase	99%			
	BBA56188.1 hypothetical protein BBTM_01911	80%	**LPGTG**	KGALP**T**V**V**K**K**	YTLT**E**T**E**APA**G**Y

Table 3.4. (Continued)

				NNNTL**TV**AM**K**	
	BBA56187.1 cell surface protein	96%	**LP**L**TG**	GADCT**TV**TQ**K**	
	BBA55962.1 sortase	98%			
	BBA55960.1 hypothetical protein	94%		NGYQF**TV**SD**K**	
Cluster 2	BBTM_01596		**L**KY**TG**		
				DTLKV**TV**DN**K**	
	BBA55959.1 fimbrial subunit protein	94%	**LP**L**TG**	IGAGV**TV**GV**K**	YTIE**EI**AAPN**G**Y
	BBA56409.1 putative von Willebrand factor type A domain-containing protein	98%	**LP**M**TG**	SDYTV**TV**SG**K**	
				DGVTY**TV**TF**K**	
				GNGSV**TV**TL**K**	
Cluster 3	BBA56410.1 fimbrial subunit protein	95%	**LP**K**TG**	VDTAA**TV**TF**K**	YTVTE**TA**VAD**G**Y
				GGAAA**TV**YA**K**	
	BBA56411.1 sortase family protein	100%			

Further the SD pili clusters were identified in 37 strains of *B. bifidum* including the complete genomes. Among which 22 strains encode for 3 pili clusters, 14 for 2 clusters and 1 strain have a single cluster. Based on their number of pili clusters, pilin motif, CWSS motif and E box sequence the strains were grouped into 4 groups (Supplementary Table 3.1). The comparative analysis shows that the SD pili shows genetic diversity in both the sequence and number of pili in each *B. bifidum* strain.

3.4. Conclusion

In this study one of the *B. bifidum* strain TMC3115 was analysed. The strain is previously found to show inhibitory effect in allergic inflammation [96]. A detailed bioinformatics analysis was carried out to compare it with other *B. bifidum* strains. This study focused on identifying its genomic content and elucidate the extracellular factors such as extracellular proteins and pili which have the role in host interaction and immunomodulation.

The investigation of genomic content of TMC3115 strain revealed the variability in the genomic structure of this strain. In TMC3115 strain an inversion of ~ 382 kb was detected around the replication origin. Large genomic rearrangements are generally not common among

the strains of same species in bacteria more specifically non-pathogenic however, it is described in some of the probiotic species like *B. breve* JCM 7017 [129] *B. longum* subsp. *infantis* strain ATCC15697[60], *Lactococcus lactis* [130] *and Lactobacillus johnsonii* [115]. It has been shown in various studies that inversions might not have a detrimental effect on its phenotype or growth [114]. Although the inversion observed in TMC3115 is disturbing the replication symmetry yet it has some major effect on its genome integrity is still not determinant. More detailed studies focusing the effect of this inversion are required to make conclusive results.

The analysis resulted in identification of 310 proteins as potential extracellular proteins in genome of TMC3115. Among the predicted extracellular proteins, those having important binding domains which have their role in host interaction were identified. Three sortase dependent pili clusters were also identified in TMC3115 genome. The proteins in these pili clusters show high similarity with sortase dependent pili cluster in PRL2010 which previously reported to show immunomodulatory activity [51]. The comparative analysis of SD pili clusters among the *B. bifidum* strains revealed the diversity in the gene number and sequence for pili encoding. This correspond to potential variability among the *B. bifidum* strains in their adhesion to mucosal walls and host interaction.

Overall the study reveals the genomic features and structure of TMC3115 strain in detail. Moreover, it provides insight into the extracellular structures which might have their role in host interaction and immunomodulation.

CHAPTER 4

Comparative Genomic Analysis of *Bifidobacterium* species Isolated from Egyptian Fruit Bat *Rousettus aegyptiacus*

4.1. Introduction

Bats are geographically prevalent except for the Antarctica and genetically diverse. They also vary in size from the largest golden-capped fruitbat (*Acerodon jubatus*), with weight of about 1 kg, to the smallest, 2-g bumblebee bat (*Craseonycteris thonglongyai*) [131]. Bats play an important role in ecosystem but little is known about their gut-microbes [132]. Diet is a major factor in shaping the type of gut microbes [133]. Most bat species in the order *Chiroptera* intake diverse diet such as insects, small mammals, fish, blood, nectar, fruit, and pollen [131]. The Egyptian rousette bat (*Rousettus aegyptiacus*) in the order *Chiropetra*, on the other hand, is frugivorous, i.e., consuming only the pulp and juice of variety of fruits [134]. Their distribution is wide: from North Africa, Egypt, Cyprus, south of Turkey, eastern part of Arabian Peninsula and eastern Pakistan and northwest India. In drier regions, they eat mostly dates and fig.

In mammals, neonates nourish only on milk. Bats also depend on milk until they grow up to 70% of adult size and it is twice the average size of other mammals at weaning (37%) [135]. Among the microbial communities residing in infants in different mammals, *Bifidobacterium* is the dominant bacterial group [136,3]. Various studies have shown that *Bifidobacterium* impart beneficial health effects on their hosts such as immune modulation, prevention of pathogenic attachment and alleviation of atopic dermatitis and allergies [137,138]. Bifidobacteria are generally host-animal specific and can be separated into human type, animal type and insect type. Various studies have proposed the importance of *Bifidobacterium* species isolated from humans and different animals, like rodents, bovine rumens, rabbit and pig. Modesto *et al*. in a recent study have isolated two novel *Bifidobacterium* species from Egyptian fruit bat [11], in addition to four known species (*Bifidobacterium callitrichos, Bifidobacterium tissieri, Bifidobacterium myosotis* and *B. reuteri*). The host diet contributes to the development of intestinal microbial communities,

and the bat dietary habits should affect the development of important probiotic bacterial species like bifidobacteria.

The aim of this study was to investigate the genetic biodiversity of bifidobacteria from bat compared to bifidobacterial species from human and non-human primates by decoding genome sequences. The description of the genomic features in different niches (bat, non-human primates and human being) is fundamental in clarifying repertoire of genes that have caused their evolutionary differentiation. Such genomic analyses support the hypothesis that bat strains have been subjected to genetic adaptations to their host environment such as a peculiar diet heavily based on sugars.

4.2. Methods

4.2.1. *General feature prediction*

Genomes of 8 bifidobacterial strains from bat were isolated, sequenced, assembled and annotated as previously described [11]. Total 75 bifidobacterial strains were collected from NCBI Assembly Database (Additional file 1: Table S1) and were annotated using DFAST web server [76]. Orthologous clustering for was conducted using GET_HOMOLOGUES software [74]. The parameters for the orthologous clustering were as follows: E-value threshold of 10e-5, minimum percentage coverage of 75%, and the algorithm, OrthoMCL.

4.2.2. *Pan and core genome determination*

Cluster of Orthologous Group (COG) functional categories were used to identify the pan and core genome. Reverse Position-Specific BLAST was used to query orthologous gene clusters against the NCBI-CDD and COG categories were assigned using Perl script "cdd2cog" (https://github.com/aleimba/bac-genomics-scripts/tree/master/cdd2cog). Pan genome is selected as orthologous clusters present in the genomes and core genome was selected as described previously [41]. Unique proteins for bat strains were assigned Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology using Blast-KOALA [139].

4.2.3. *Prediction of carbohydrate-active enzymes and transport systems*

Carbohydrate active enzymes (CAZymes) for all the bifidobacterial strains were identified using Carbohydrate Active Enzymes (CAZy) database. For annotation of carbohydrate-active enzyme, dbCAN online web server was used [79]. Carbohydrate transport proteins were predicted using Transporter Classification Database (TCDB) [140].

4.2.4. *Statistical Analysis*

To compare the glycosyl hydrolases (GHs) among different groups with non-equal sample size Kruskal-Wallis test with significance level of $p > 0.05$ was done. Further to check which groups are significantly different Dunn's post hoc test was run. All these analyses were performed using R version 3.6.2.

4.2.5. *Phylogenetic Analysis*

The strict core protein sequences (355) identified by orthologous clustering were aligned using MAFFT program (version 7.313) [141]. The alignments were trimmed using trimAl [142]. The phylogenetic tree was built using RaxML (version 8.2.7) with PROTGAMMA-BLOSUM62 substitution model and 1000 rapid Bootstrap searches [143].

4.3. Results and Discussion

4.3.1. *General characteristics of bifidobacterial genomes from bat*

Eight bifidobacterial strains were isolated from Egyptian rousette bat and their genomes were determined (Table 4.1).

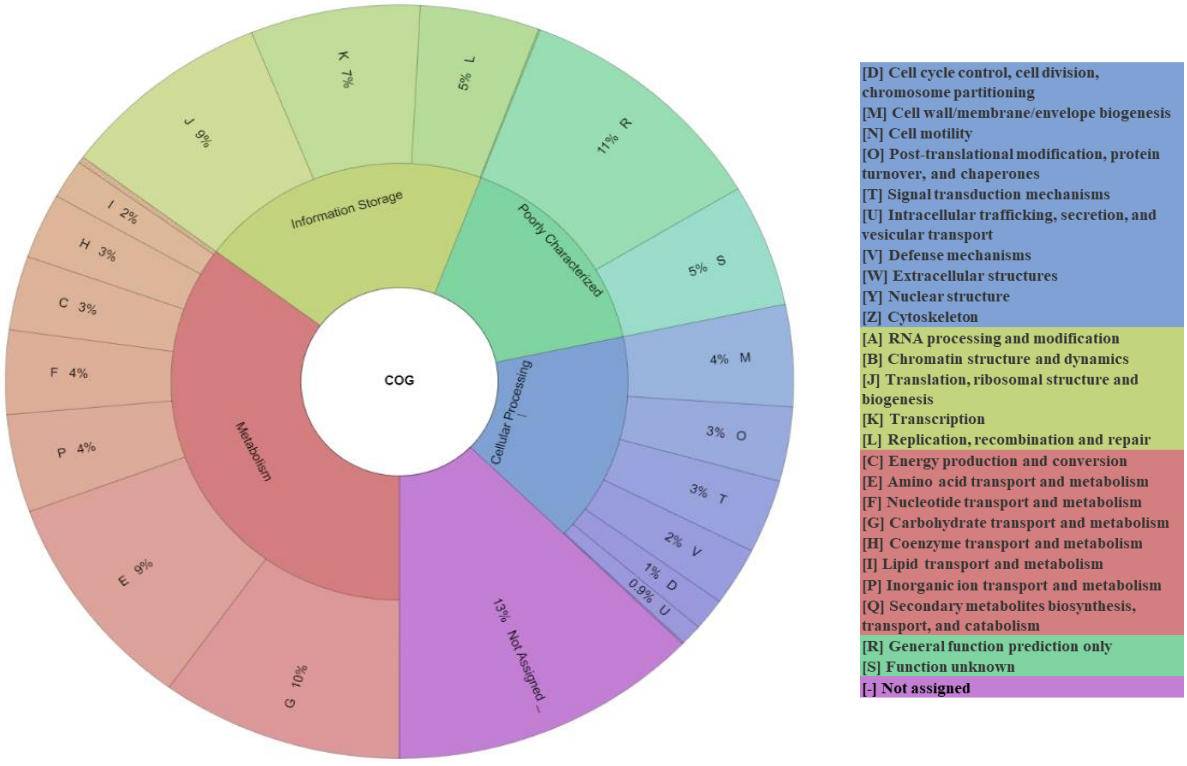
Table 4.1. General genomic characteristics of bat-isolated bifidobacterial species

Species	Genome length	Number of Genes	GC Content	Number of rRNAs	Number of tRNAs	Number of CRISPRs	IS elements/ transposases
***B. vespertilionis* (strain 1)**	3075992	2409	64.2	1	62	5	14
***B. vespertilionis* (strain 2)**	3067389	2406	64.2	1	64	5	7
***B. rousetti* (strain 3)**	3053799	2593	64.6	1	68	7	14
***B. tissieri* (strain 4)**	3032244	2385	61	1	63	3	15
***B. tissieri* (strain 5)**	2986510	2481	60.8	2	63	4	10
***B. myosotis* (strain 6)**	3275217	2575	63.2	0	67	7	13
***B. reuteri* (strain 7)**	2833112	2239	60.4	0	59	4	19
***B. callitrichos* (strain 8)**	2797830	2264	63.6	3	64	1	12

Three were identified as new species (two *Bifidobacterium vespertilionis* and one *B. rousetti*) and the remaining 5 belonged to known species from non-human primates (two *B. tissieri, B. myosotis*, *B. reuteri*, and *B. callitrichos*). Their genome size ranged from 2.8 to 3.28 Mb. The G+C content ranged from 60.4 to 64.6 %. The new species possess larger genome size and higher G+C content compared to the others.

The number of core gene families among the 8 strains was 1552 [41]. Their COG (Cluster of Orthologous Groups) distribution ranged as follows: metabolism related (35%), information storage and processing (22%), poorly characterized (16%), cellular processing and signalling (14%), and unassigned (13%) (Figure 4.1a). The distribution of each COG category was almost similar for all strains (Figure 4.1b). The pan genome (all genes) contained about 24,000 gene families, among which 1487 were bat-specific. Among them, 78% genes were without COG category. Excluding the unassigned categories, 15% - G (Carbohydrate transport and metabolism), 13 - were S (function unknown), 13% - K (Transcription) ,11% - L (Replication, recombination and repair) and others were less than 10% (Supplementary Figure 4.1a and b).

a



b

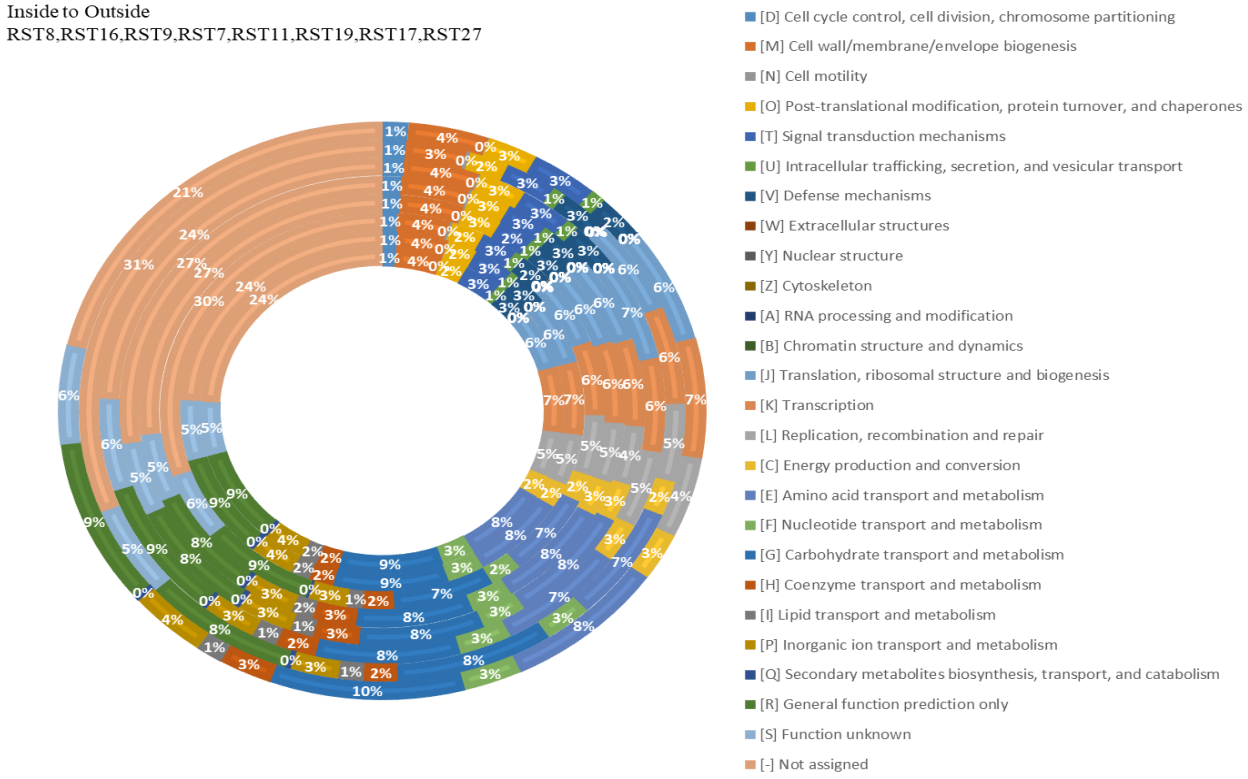


Figure 4.1 (a) COG statistic of the core genome of bifidobacterial species from bat, Korna chart: KornaTools v2.7 [24] (**b)** COG categories distribution in all bat isolated bifidobacterial species.

Only 108 (7.3%) of the unique genes were assigned KEGG functional categories. Among those with assigned categories, mostly the proteins related to Protein families: signaling and cellular processes (ko02000-Transporters, ko02048-Prokaryotic defense system), Environmental Information Processing (ko02010- ABC transporters), Genetic Information Processing (ko03060-Protein export), Protein families: genetic information processing (ko03400-DNA repair and recombination proteins), Carbohydrate metabolism (ko00562 Inositol phosphate metabolism, ko00500 Starch and sucrose metabolism, ko00520 Amino sugar and nucleotide sugar metabolism) were present (Figure 4.2).

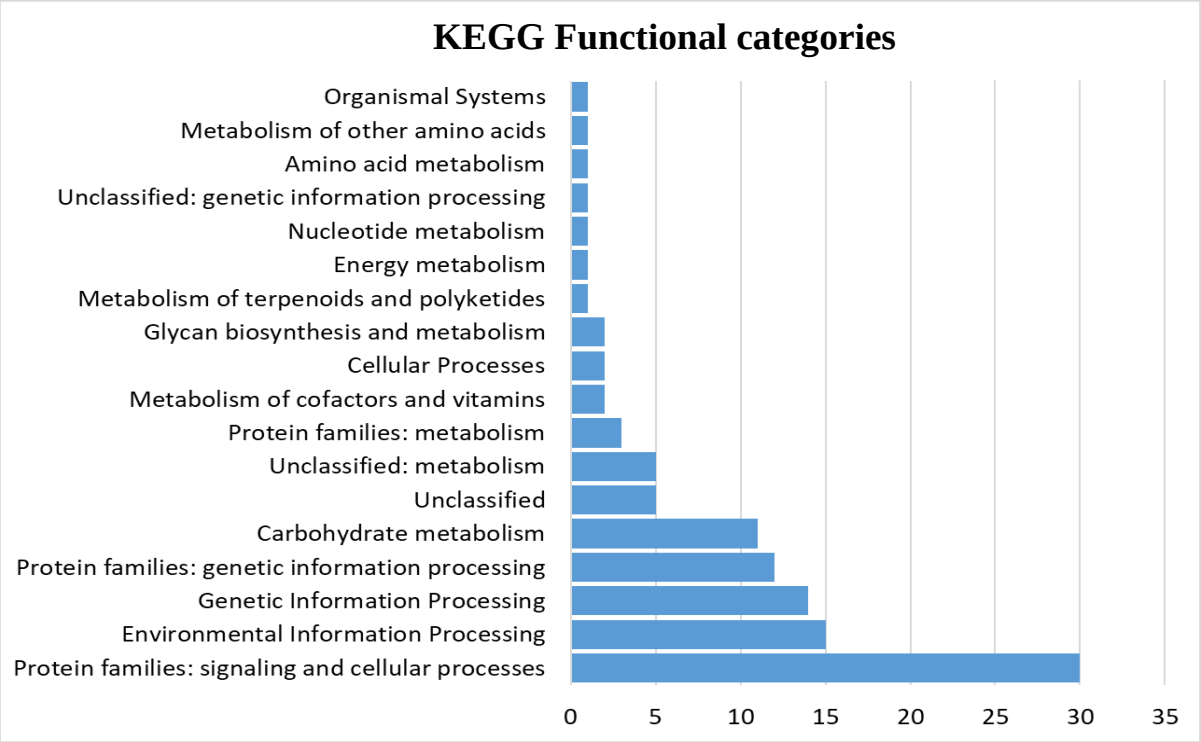


Figure 4.2. Distribution of KEGG functional categories for unique genes. The x-axis shows the number of corresponding proteins in functional category, and the y-axis shows the KEGG functional categories.

Detail analysis of unique genes showed the presence of specific gene cluster of ABC transporters involved in maltose/maltodextrins utilization. Maltose uptake system ($MalFGK_2$-E) in *E. coli* and *Salmonella Typhimurium*, is composed of a periplasmic maltose-binding protein (MalE), two integral membrane proteins (MalF and MalG), and two copies of the cytoplasmic ATP-binding cassette (MalK). Previous studies show that, in *Bifidobacterium* species operon for maltose transport contains malEFG without ATPase, which is present as a standalone conserved gene and is not co-regulated with malEFG [144]. However, in *B. vespertilionis* (strain 1, 2), there was a unique operon with a maltose-binding protein (malE), and two membrane spanning ABC transporters (malF and malG), and an ATP binding protein belonging to sugar ABC transporter family (malK) (Figure 4.3a, 4b). The operon with MalEFGK proteins shows higher similarity of amino acids with those in Firmicutes, suggesting a different evolutionary origin.

Other unique genes included two genes for alpha-L-fucosidase (GH29) and one for beta-galactosidase with LacZ domain. The alpha-L-fucosidase genes possess only the Pfam01120 alpha-fucosidase domain and no N-terminal signal sequence for secretion, suggesting it as intracellular similar to the α-L-fucosidases found in *B. longum* and *Lactobacillus casei*. However, in the habitats where fucosyloligosaccharides presume to be important energy and carbon source like in *B. bifidum* they are extracellular [60,145]. Such a variability in genes suggest the bacterial host adaptability. Addition of these genes reflect the unique metabolic ability of bat *Bifidobacterium*.

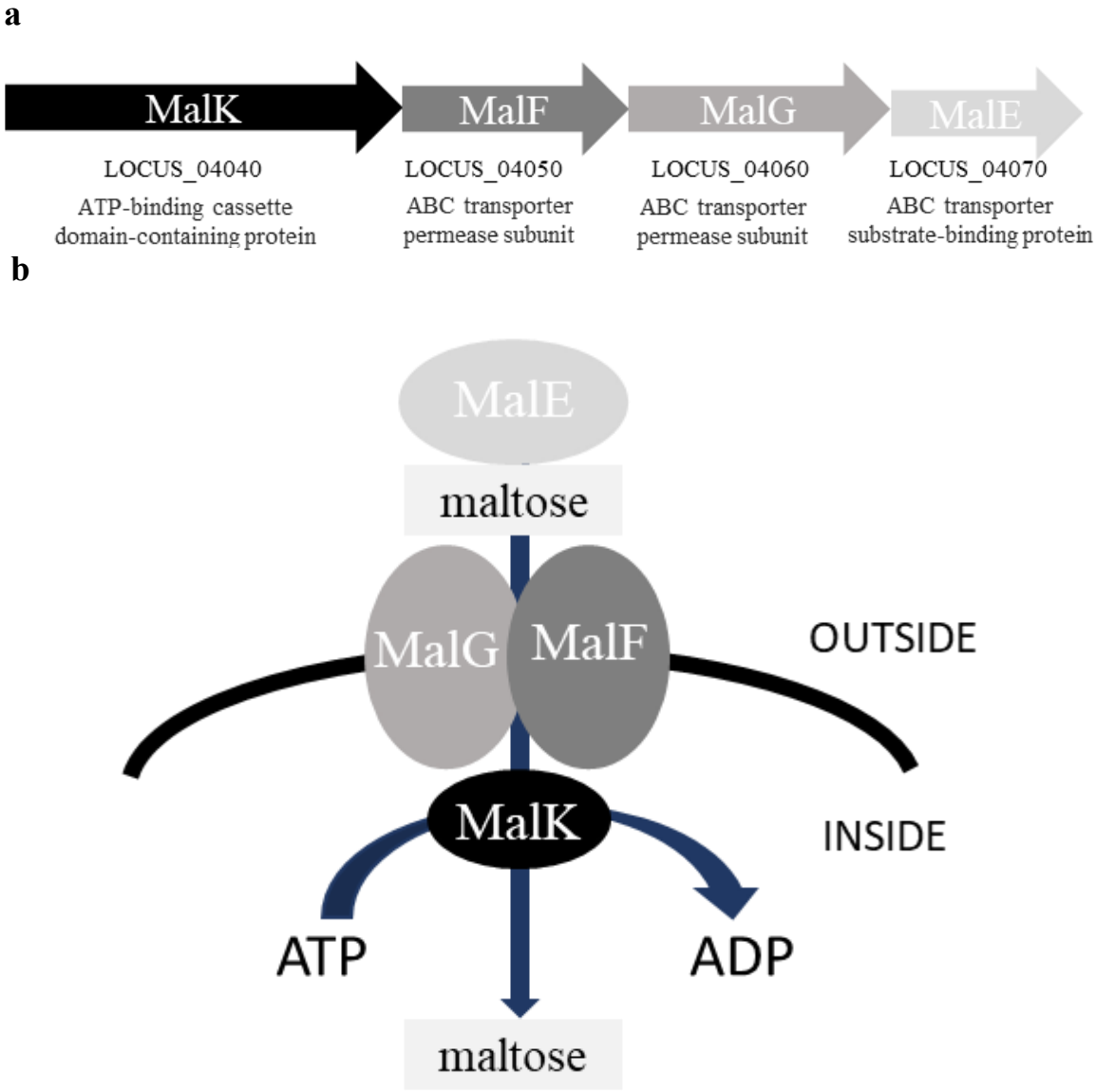


Figure 4.3. (a) MalEFGK operon in *B. vespertilionis* (strain 1,2). Each arrow represents a gene. The length of the arrow is proportional to the predicted gene size. Each gene is marked with different colour. The locus and putative function of each gene is indicated below the arrow. **(b)** Transport of maltose in *B. vespertilionis* (strain 1,2). The ABC transporters encoded by the operon (MalEFGK) are shown.

Table 4.2. Average number of GHs involved in milk oligosaccharide metabolism among the bat and human infant group

	GH 2	GH 20	GH 29	GH 33	GH 42	GH 95	GH 112	GH 136
Bat	6.625	1.375	0.625	0	4.25	0	0.625	0.125
Human Infant	3	1.25	0.875	0.75	2.5	0.625	0.5	0.125

4.3.2. *Carbohydrate utilization by bat isolates*

In the bat isolates the most abundant COG category was carbohydrate transport and metabolism. On average 10% of the bat bifidobacterial genes were involved in carbohydrate transport and metabolism.

4.3.2.1. *Milk oligosaccharides*

Among mammals' bat is unique in their foraging behavior. The bats don't start foraging until grown nearly up to adult size. They depend highly on milk for their growth [135]. Comparison of milk oligosaccharides from bats with different dietary habits like insectivorous and frugivorous suggested that diet contribute differences in milk composition. The milk of fruit and nectar eating bats contains more lactose compare to that of insectivorous [146,147]. The comparison of milk oligosaccharides in different mammals by Urashima et al showed that no fucosyl oligosaccharides were detected in bat milk [148].

Studies of bifidobacteria from human infants and calves have elucidated the role of bifidobacteria in metabolizing human and bovine milk oligosaccharides, respectively [149,150]. It is probable that bifidobacteria from bat also have a role in bat milk metabolism. The comparison of GHs involved in milk oligosaccharide metabolism among the bat and human infant isolated bifidobacterial species reveal that the Egyptian fruit bat isolates have high number of GH2 (β-galactosidase) and GH42 (β-galactosidase) and have relatively less genes for GH 29 (α-Fucosidase) and no genes for GH 95 (α-Fucosidase) (Table 4.2). Gain of more genes for lactose metabolism and loss of genes for fucose metabolism in the bat isolates suggest adaptation to their host according to their dietary pattern.

4.3.2.2. *Sugar metabolism*

The evaluation of gene clusters for sugar metabolism in all the six bat strains was done using homology search with the known genes for sugar metabolism. The analysis revealed that all of these strains have gene clusters for the metabolism of raffinose and fructo-oligosaccharides while for other sugars like xylose, ribose, fructose, maltotriose, sucrose and lactose there was variability (Figure 4.4).

Figure 4.4. Different sugar metabolism genes in bat bifidobacterial species. Green colour shows the presence of genes based on the amino acid similarity while red colour shows the absence of gene.

4.3.3. *Carbohydrate transport systems*

Genes encoding the carbohydrate transporters in all the strains were predicted based on Transporter Classification Database. Table 4.3 shows the number of carbohydrate transporters belonging to six different groups of ABC-type family, PEP-PTS systems, major intrinsic protein (MIP), major facilitator superfamily (MFS), glycoside-pentoside-hexuronide (GPH): cation symporter family, and glucose/ribose porter family (GRP). *B. vespertilionis* (strain 1 & 2), *B. reuteri* (strain 7) and *B. callitrichos* (strain 8) had more than 100 genes for carbohydrate transporters. *B. tissieri* (strain 4 & 5) had the least number of carbohydrate transporters.

The ABC transporter systems are involved in transport of ribose, maltose, lactose, FOS, alpha-glucosides, raffinose, mannose and xylose while PEP-PTS systems transport glucose using glucose-specific PEP-PTS. The homologues of ABC transporter genes for ribose,

maltose, lactose, raffinose, xylose and FOS of *B. longum* NCC2705 were identified in different strains. Strain 3 and 8 contained complete PTS systems as the *B. bifidum* PRL2010 strain. They had general components of this system histidine protein (HPr) and enzyme I (EI), and also the variable components EIIA, EIIB, and EIIC present. All the other strains only had the general component of PTS system HPr and EI (Supplementary Figure 4.2).

Table 4.3. Carbohydrate transport systems of bat-isolated bifidobacterial species

	ABC	**PTS**	**MFS**	**GPH**	**GRP**	**MIP**	**Total**
***B. vespertilionis* (1,2)**	61	3	30	10	4	3	111
***B. rousetti* (3)**	54	6	25	6	3	2	96
***B. tissieri* (4,5)**	23	2	33	14	3	2	77
***B. myosotis* (6)**	71	2	27	8	4	2	114
***B. reuteri* (7)**	50	2	27	8	4	2	93
***B. callitrichos* (8)**	63	3	30	5	4	3	108

*number(s) in the round bracket represent the strain

One of the studies on sugar transport systems in *B. longum* species suggest that the role of PTS in bifidobacterial species varies. Some species have glucose and some have fructose PTS system [65]. Maze *et al*. have reported a fructose PTS in *B. breve* genome [64]. This PTS system is similar to that of *B. longum* with one gene glcP which encodes a glucose/proton symporter is missing. PTS system of the bat species were analysed in detail and the results showed that species in three of the cluster also lack gene glcP as in *B. breve* and show high similarity to fruA gene (encodes EIIBCA). These results suggest that the PTS system in strain *B. rousetti* (strain 3), *B. tissieri* (strain 4 & 5) and *B. callitrichos* (strain 8) where the glcP gene is missing might have fructose-PTS while *B. vespertilionis* (strain 1 & 2), *B. myosotis* (strain 6) and *B. reuteri* (strain 7) have glucose-PTS system (Table 4.4, Figure 4.5).

Table 4.4. Comparison with *B. breve* fruA and *B. longum* ptsG and glcP genes: Values show the amino acid identity, expressed in percentages.

	ptsG	***fruA***	***glcP***
***B. vespertilionis* (1,2)**	64.98	53.69	Present
***B. rousetti* (3)**	81.36	59.09	Absent
***B. tissieri* (4,5)**	48.88	54.21	Absent
***B. myosotis* (6)**	51.8	55.57	Present
***B. reuteri* (7)**	51.62	56.61	Present
***B. callitrichos* (8)**	78.98	52.94	Absent

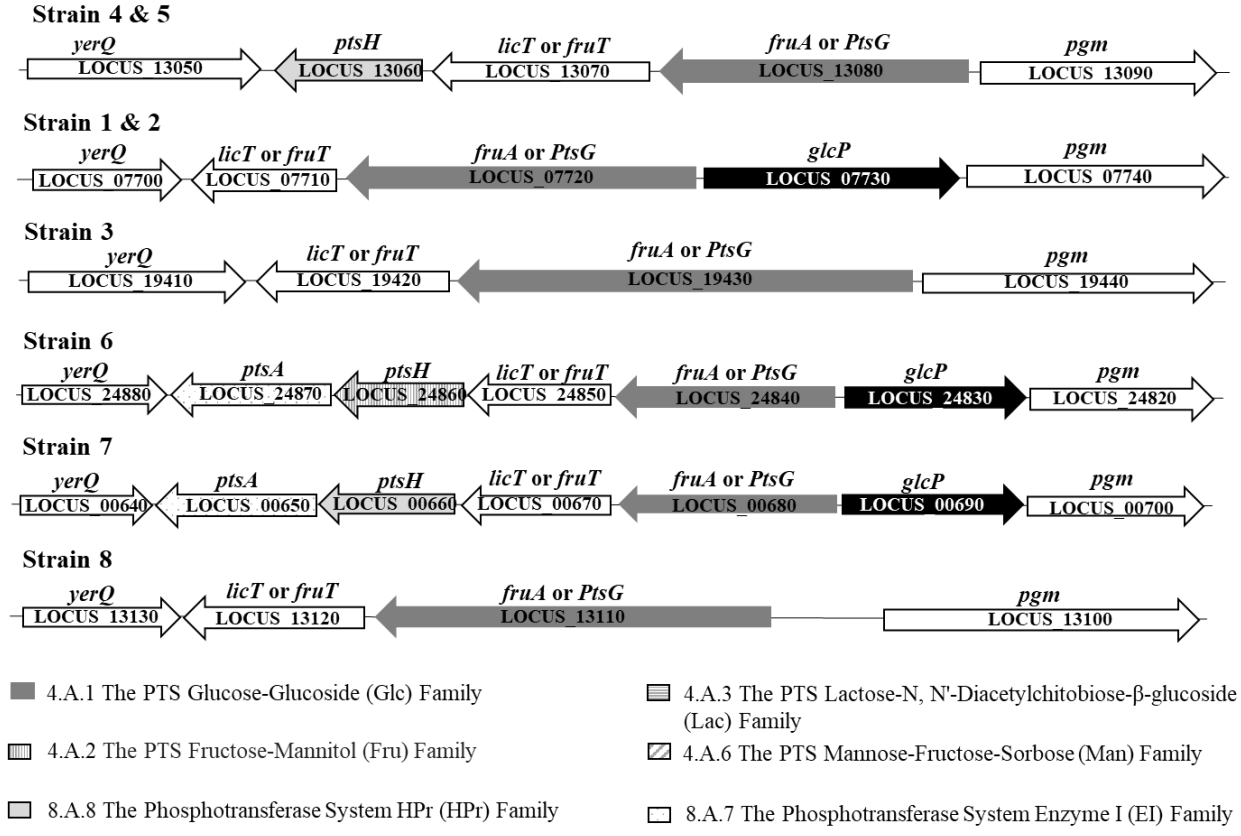


Figure 4.5. Comparison of gene cluster encoding homologues of FruA and PtsG in *B. breve* and *B. longum* respectively, *fruA* and *ptsG (*EIIBCA), *licT* and *fru (*transcriptional antiterminator), *pgm (*phosphoglucomutase), *glcP (*glucose symporter) , *yerQ (*sphingosine kinase), ptsH (phosphocarrier protein HPr), ptsA(phosphoenolpyruvate-protein phosphotransferase). Genes are shown by arrow and color marks the type of PTS.

4.3.4. *Comparative analysis of bat isolates with other Bifidobacterium species*

To evaluate the similarities of bat strains with other Bifidobacterium strains, the core genes of bat strains were compared with the Bifidobacterium species grouped according to their isolation sources. The results revealed that bat strains shared the highest percentage of its core genes with non-human primate species (Table 4.5). Further the phylogenetic analysis based on core genes also showed that bat strains are mostly clustered with strains isolated from non-human primates (Figure 4.6). This propose similar genetic abilities of bat and non-human primate species.

Table 4.5. Percentage of shared genes with bat core genes

	Non-human primates	**human adult**	**human infant**	**Fermented Products**	**Insects**	**Rodent**
Bat	94 %	88 %	87 %	82 %	74 %	82 %
	Bovine Rumen	**Rabbit**	**Pig**	**Birds**	**Sewage**	**Commercial Products**
Bat	84%	82%	84%	84%	82%	83%

Further to evaluate the carbohydrate metabolism, CAZymes in all the strains were identified. Comparison of the CAZymes among the four groups bats, non-human primates, humans and others show that species from bats and non-human primates have comparatively high number of CAZymes (Figure 4.7). The analysis revealed that bat isolates have greater than the average number of GH and high number of β-galactosidases genes belonging to GH2, presumably involved in the utilization of milk substrates [151,152], degradation of mucin and plant polymer galactan [61,153]. The relatively high number of β-galactosidases genes in bat species indicate the importance of milk and plant galactans in the bat diet.

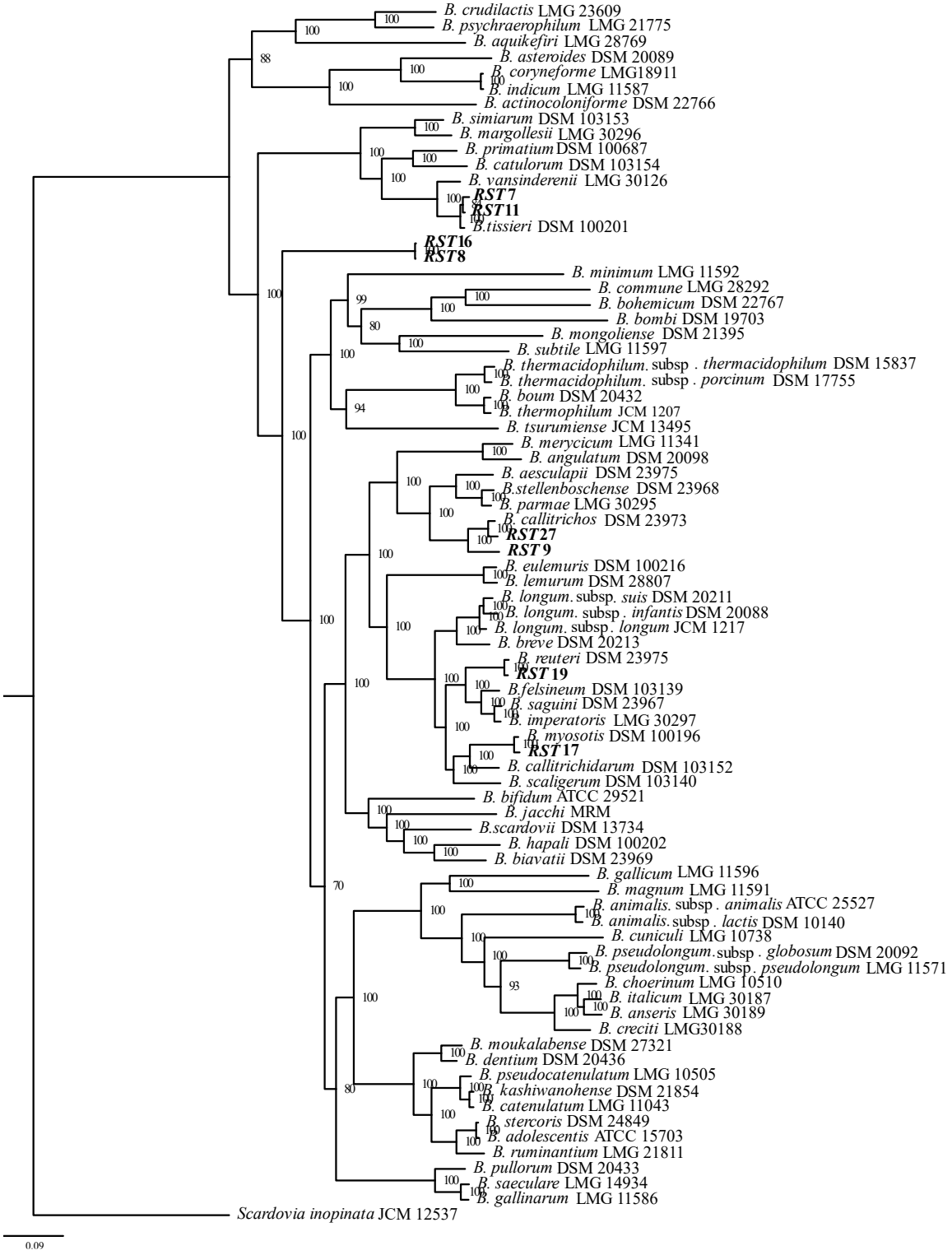


Figure 4.6. Phylogenetic tree based on amino acid sequences of 355 core genes. The strains isolated from Egyptian fruit bats are highlighted. Maximum likelihood method was used to built the tree with sequences of *Scradovia inopinata* used as an outgroup.

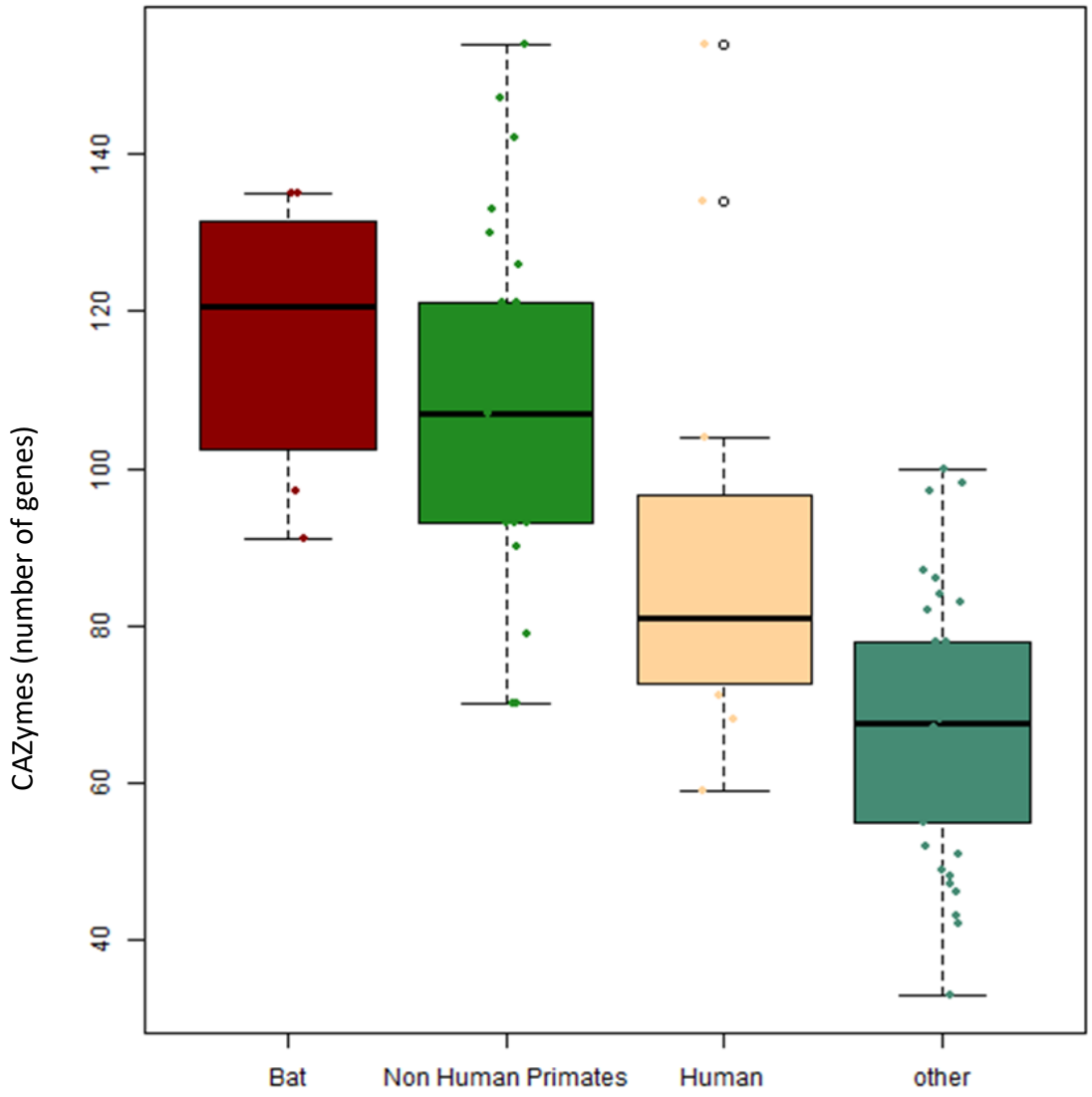


Figure 4.7. Distribution of CAZymes among the species grouped into four groups bats, non-human primates and others. The circles show the data points.

Statistical comparison of GHs shows that GH88 (d-4,5-unsaturated β-glucuronyl hydrolase) is significantly higher in bat group. GHs classes, i.e. GH2 (β-galactosidases), GH 59 (β-galactosidase), GH 78 (α-L-rhamnosidase), GH 105 (xylan α-1,2-glucuronidase) and GH 115 (unsaturated rhamnogalacturonyl hydrolase) are significantly higher in bat and non-human primate species (Table 6). These GHs specific to bat and non-human primate species have their role in metabolizing plant-derived carbohydrates (e.g., pectin, hemicellulose, and xylans).

The presence of these important plant metabolizing GHs proposes the dietary relationship between these groups.

Table 4.6. Post hoc comparison using Dunn's test among glycosyl hydrolases (GHs) classes in different groups. Comparison of bat with all groups. The highlighted cells show that these categories are significantly different from bat group (p-value< 0.05).

GHs	**Groups**	**Dunn test p-values adjusted**
GH 88	Bat - Humans	0.00157138
	Bat - Non-Human Primates	0.047674822
	Bat - Others	0.000141495
GH 2	Bat - Humans	3.76E-02
	Bat - Non-Human Primates	1.47E-01
	Bat - Others	8.68E-07
GH 59	Bat - Humans	9.15E-04
	Bat - Non-Human Primates	5.62E-02
	Bat - Others	9.34E-06
GH 78	Bat - Humans	0.00781941
	Bat - Non-Human Primates	1
	Bat - Others	0.003008142
GH 105	Bat - Humans	0.008600629
	Bat - Non-Human Primates	0.645967271
	Bat - Others	0.002891912
GH 115	Bat - Humans	1.38E-03
	Bat - Non-Human Primates	6.22E-01
	Bat - Others	4.43E-05

Further the GHs between the same species isolated from bat and non-human primates were compared. In *B. reuteri* from bats none of the gene for class GH 43 (β-xylosidase, α-L-arabinofuranosidase) are present while in *B. reuteri* from non-human primates seven genes for this class are present. Such differences reveal that specific gene set such as genes having role in carbohydrate metabolism plays role in bifidobacterial host adaptation.

4.4. Conclusions

In this study the role of bifidobacterial species specific to bat was examined. The comparative analysis revealed that the strains from bat possess a high percentage of carbohydrate transport and metabolism genes, including the high number of β-galactosidases genes belonging to GH2 possibly for digestion of milk and other plant carbohydrates.

These strains share genomic and specifically metabolic similarities with non-human primate species than the other mammalian species or human. These metabolic similarities probably reflect the food categories of their host, i.e., various fruits are also edible for non-human primates but are different from forage of ruminants.

Even within the species from bat, there is a variability in the sugar and carbohydrate metabolism. The species i.e. *B. vespertilionis* (strain 1 & 2) have more genes involved in milk metabolism, *B. myosotis* (strain 6) shows the genes for metabolism of several sugars, *B. vespertilionis* (strain 1 & 2) and *B. myosotis* (strain 6) seems to have better ability to utilize plant carbohydrates, *B. vespertilionis* (strain 1 & 2), *B. myosotis* (strain 6) and *B. callitrichos* (strain 8) have a wide set of carbohydrate transporters and *B. rousetti* (strain 3), *B. tissieri* (strain 4 & 5) and *B. callitrichos* (strain 8) have fructose-PTS while other species have glucose-PTS. Differences in carbohydrate metabolism and variability suggest the mutualistic characteristics of these species.

It is concluded that bifidobacterial strains from the bat contribute synergistically to the complex carbohydrate metabolism and are well adapted to the dietary habits of its host. They share metabolic similarities with other *Bifidobacterium* species in non-human primates in accordance with their dietary associations.

CHAPTER 5

Host-diet Effect on the Metabolism of *Bifidobacterium*

5.1. Introduction

Commensal gut bacteria are environment-specific and evolve together with their hosts. The genus *Bifidobacterium* is a widespread and abundant genus belonging to phylum Actinobacteria and is mainly distributed in intestinal environments of various animals, from insects to mammals [9-14]. They have been considered beneficial microorganism useful to host health status. For what concern humans, they are the first colonizers of gut microbiota; a vertical transmission from mother to offspring in humans but also in other animals plays a fundamental role in bifidobacterial occurrence in the gut microbiota. Moreover, colonization of bifidobacteria is modulated by "indigestible" carbohydrates, such as oligosaccharides derived from breastmilk in mammals and plants. These compounds together with the physiology of the host are important drivers of bifidobacterial host co-evolution. It has been shown that certain bifidobacterial species are both host- and niche-specific. Examples of host-specific species are *B. breve* for humans, *B. rousetti* for bat and *B. reuteri* for marmoset [77,27]. On the other hand, there are some species with cosmopolitan life style such as *B. longum*, isolated from humans and animals, and *B. animalis* and *B. pseudolongum* isolated from different animal species. Since whole genomes are available for many *Bifidobacterium* strains belonging to different species, several genome-scale analyses revealed the acquisition of specific genes allowing their host specificity [94].

The genomic reservoir of the genus shows an open pan-genome, harboring a large number of strain-specific genes. The genome composition of host-specific strains shows weak association with the phylogeny of their host animals, especially in terms of accessory genes for amino acid production and carbohydrate degradation [154]. Notably, bee-derived species cluster themselves in a deep branch with small genome sizes [155]. Despite multiple attempts, however, identification of host specificity and elucidation of its mechanism has remained unclear from the whole genome analyses.

This study focuses on the relationship between host diets and bacterial glycoside hydrolases (GHs) to investigate the evolutionary relationship between bifidobacteria and host animals. To identify this relationship, bifidobacterial species were classified into 13 different groups based on their host dietary patterns. A comparative analysis approach was used to inspect the genomic features such as genome size and GH gene content among the dietary groups. The phylogenetic relationship among the species was also assessed and the phylogenetic signal for the GH content was calculated. The comparative analysis provides insight into bifidobacterial adaptation to ecological niches.

5.2. Materials and Methods

5.2.1. *Genomic data and annotations*

For the genus-level classification, the type strain data of the 84 recognized *Bifidobacterium* taxa with 76 species and 8 subspecies (*Bifidobacterium animalis* subsp. *lactis*, *B. longum* subsp. *infantis*, *B. longum* subsp. *suis*, *B. catenulatum* subsp. *kashiwanoense*, *B. pseudolongum* subsp. *globosum*, *B. pullorum* subsp. *gallinarum*, *B. pullorum* subsp. *saeculare*, *B. thermacidophilum* subsp. *thermacidophilum*) were used (Supplementary Table 5.1). For the multi-host analysis, 66 strains from hosts with varying feeding behavior were used (Supplementary Table 5.2). For the analysis on *B. animalis* subsp. *lactis*, 45 strains were used (Supplementary Table 5.3).

Genomic sequences were collected from the NCBI Assembly Database and annotated by the DFAST stand-alone software program [156]. Cluster of Orthologous Group (COG) functional annotations were assigned by performing the Reverse Position-Specific BLAST against the NCBI-CDD and by the Perl script "cdd2cog" (https://github.com/aleimba/bac-genomics-scripts/tree/master/cdd2cog). The host and diet information for each strain was collected manually from the NCBI databases and related publications.

5.2.2. *Orthologous gene clustering*

Orthologous gene clustering was performed using the GET_HOMOLOGUES software package [74] (cutoff: E-value 1.0×10^{-5}, with minimum percentage coverage of 75%) and clusters were detected by the OrthoMCL algorithm [75]. Gene clusters constituting the pan-genome and the core-genome were selected based on the trend of the COG categories. The ratios of COG classes

among different set of core genomes (from 100% to 83% core) was compared and an appropriate core was chosen [41].

5.2.3. *Identification of carbohydrate-active enzymes*

The HMMER search against the dbCAN2 HMM database was used to determine carbohydrate active enzymes (CAZymes) [157]. The definition of GH families also follows the CAZy database. The standalone version of dbCAN annotation tool was used to determine their annotations.

5.2.4. *Selection of the GH families for clustering Bifidobacterium strains*

To classify the bacterial strains with their GH distribution, the selection of the GH families is crucial. GH genes are non-essential, and only two families were shared by all the strains, GH3 and GH36 (Table 5.1).

On the other hand, out of 72 GH families, 24 families were present in fewer than 5 strains (< 5%). To select the GH families that are moderately shared among the strains, we created GH sets that were shared by 100% of 84 taxa, >95% of the taxa, >90%, >85%, and so on (21 sets). Based on each GH set, we performed a hierarchical clustering of bacterial taxa using the distribution of corresponding GH genes and compared results. The GH set of sharing level >20% (Set 17 Table 5.1) produced the same clustering result as >15% and >10% (Set 18 and Set 19 in Table 5.1) indicating that the classification using 32~42 GH families was stable. Therefore, I selected the threshold of >20% in this analysis.

5.2.5. *Phylogenetic analysis*

To infer the phylogenetic relationship among the type strains, the phylogenetic tree based on 362 strict-core proteins was used. The protein alignments were trimmed using trimAL (-automated 1 option) before concatenation [142], and the alignment was constructed using MAFFT version 7.313 [141]. The tree was built using RaxML version 8.27 using PROTGAMMA-BLOSUM62 substitution model and maximum likelihood method [143]. The tree was rooted with *Scardovia inopinata* JCM 12537^{T}. The statistical reliability was evaluated by bootstrap analysis of 1,000 replicates with the Bootstrap rapid hill climbing algorithm. The tree was visualized using iTOL (https://itol.embl.de/) [158].

Table 5.1. Selection of GH families for clustering. The chosen set is shown in bold [46].

GH Subsets	Sharing % in 84 taxa	Total GH families	Number of added families	Added families
Set 1	100	2		**GH3, GH36**
Set 2	95	3	1	**GH13**
Set 3	90	5	2	**GH32, GH77**
Set 4	85	8	3	**GH2, GH25, GH42**
Set 5	80	10	2	**GH31, GH43**
Set 6	75	11	1	**GH51**
Set 7	70	12	1	**GH1**
Set 8	65	14	2	**GH5, GH30**
Set 9	60	14	0	
Set 10	55	15	1	**GH127**
Set 11	50	16	1	**GH20**
Set 12	45	17	1	**GH29**
Set 13	40	19	2	**GH78, GH112**
Set 14	35	19	0	
Set 15	30	22	3	**GH38, GH120, GH136**
Set 16	25	26	4	**GH94, GH115, GH125, GH146**
Set 17	**20**	**32**	**6**	**GH95, GH129, GH59, GH26, GH35, GH28**
Set 18	15	37	5	GH109, GH105, GH33, GH8, GH27
Set 19	10	42	5	GH101, GH121, GH23, GH53, GH65
Set 20	5	48	6	GH10, GH123, GH130, GH39, GH85, GH88
Set 21	0	72	24	GH106, GH110, GH113, GH140, GH141, GH142, GH151, GH154, GH16, GH18, GH4, GH49, GH50, GH55, GH63, GH67, GH73, GH79, GH76, GH84, GH89, GH91, GH92, GH93

5.2.6. *Statistical analysis*

Kruskal-Wallis test (significance level of $p < 0.05$) and Dunn's post-hoc test was performed using the R version 3.6.2. Phylogenetic signal for genomic trait of GH content was calculated using the R package "phylosignal" [159]. GH content is defined as the percentage of GH genes in each bifidobacterial type strain. To measure the strength of the phylogenetic signal (likelihood of shared evolutionary history), we used Blomberg's K statistic [160]. The K values closer to 1 and 0 indicate strong and weak evolutionary correlation, respectively. To detect the hotspots of autocorrelation, local Moran's I for each species and local indicator of phylogenetic association (LIPA) were computed.

5.3. Results and Discussion

5.3.1. *Host diet and the genome size of type strains*

The genomic sequences for 84 *Bifidobacterium* type strains (76 species and 8 subspecies) were investigated. The genome size of the strains ranged from 1.63 to 3.25 Mb with an average of 2.43 Mb (SD ±0.40). The GC content ranged from 50.4 to 66.6% with an average of 60.8%. The orthologous clustering of their coding genes revealed that the pan-genome amounted to 24,181 gene clusters including singletons.

The number of clusters shared across ≥80 strains and across all strains were 722 and 362, respectively. The latter strict core was used to construct the phylogenetic tree by concatenating the amino acid sequences of the strict-core genes. In the resulting tree, 10 previously described groups [16] and one additional group were identified. The new group consisted of *Bifidobacterium avesanii* and *B. vespertilionis* (Figure 5.1). The former strain was isolated from cotton-top tamarin (*Saguinus oedipus*), a new-world monkey in South America feeding mainly on fruits and insects [161]. The latter, *B. vespertilionis*, was isolated from Egyptian fruit-bat (*Rousettus aegyptiacus*) feeding only on the pulp and juice of various fruits [134]. Two strains, *Bifidobacterium tsurumiense* and *Bifidobacterium minimum*, were not included in any cluster.

To examine the relationship between host diets and the genome sizes, the strains were classified into 13 dietary groups according to the feeding behavior and isolation sources of their hosts (Supplementary Figure 5.1 and Supplementary Table 5.1). Genome sizes differed

significantly among the different dietary groups (Kruskal-Wallis chi-squared = 59.101, df = 13, p-value = 7.603e-08) (Figure 5.2).

Strains from bees showed the smallest genome sizes as previously reported [11]. The genome sizes of strains from herbivores and granivores were similar. Within primate origins, the genome sizes differed between human adults and pigs, feeding on both of plant and animal matter, and monkeys feeding on fruits (frugivore), plant exudates (exudativore), or gums (gummivore). The latter showed a larger genome size while that of human and pig strains is comparable to the size in herbivores (leafs) and granivores (grains). Strains from human infants exhibited an intermediate genome size. In all groups, no significant differences were found in the GC content (Supplementary Table 5.1).

5.3.2. *Distributionution of carbohydrate-active enzymes*

The largest dietary difference between human adults and infants is milk oligosaccharides. Human milk contains diverse non-digestible oligosaccharides, classified into 13 structure series. As we shall see, GH33 (sialidase) is enriched only among strains from human infants, because sialic acid is a characteristic sugar in human milk. To investigate such metabolic correlation comprehensively, all carbohydrate-related genes were first investigated.

According to the Carbohydrate Active Enzymes (CAZy) system, each strain possessed from 33 to 166 genes (mean 88; SD ±29.46). These genes spanned the wide range of CAZy families: 72 GHs (glycoside hydrolases), 17 GTs (glycosyltransferases), 10 CEs (carbohydrate esterases) and 2 PLs (polysaccharide lyases) and 20 CBMs (appended non-catalytic carbohydrate-binding modules). Shared among ≥80% of the strains were 10 GH families (GH2, GH3, GH13, GH25, GH31, GH32, GH36, GH42, GH43, GH77), 5 GT families (GT2, GT4, GT28, GT35, GT51), CE10, and CBM48. Among these families, the distribution significantly differed ($p < 0.01$) among hosts of different diets in 7 GH families (GH2, GH3, GH13, GH31, GH36, GH43 and GH77), 3 GT families (GT2, GT4, GT35), CE10, and CBM48 (Figure 5.3, Supplementary Figure 5.2 and Supplementary Figure 5.3). Considering the diversity of the gene distribution, I focused on the GH families.

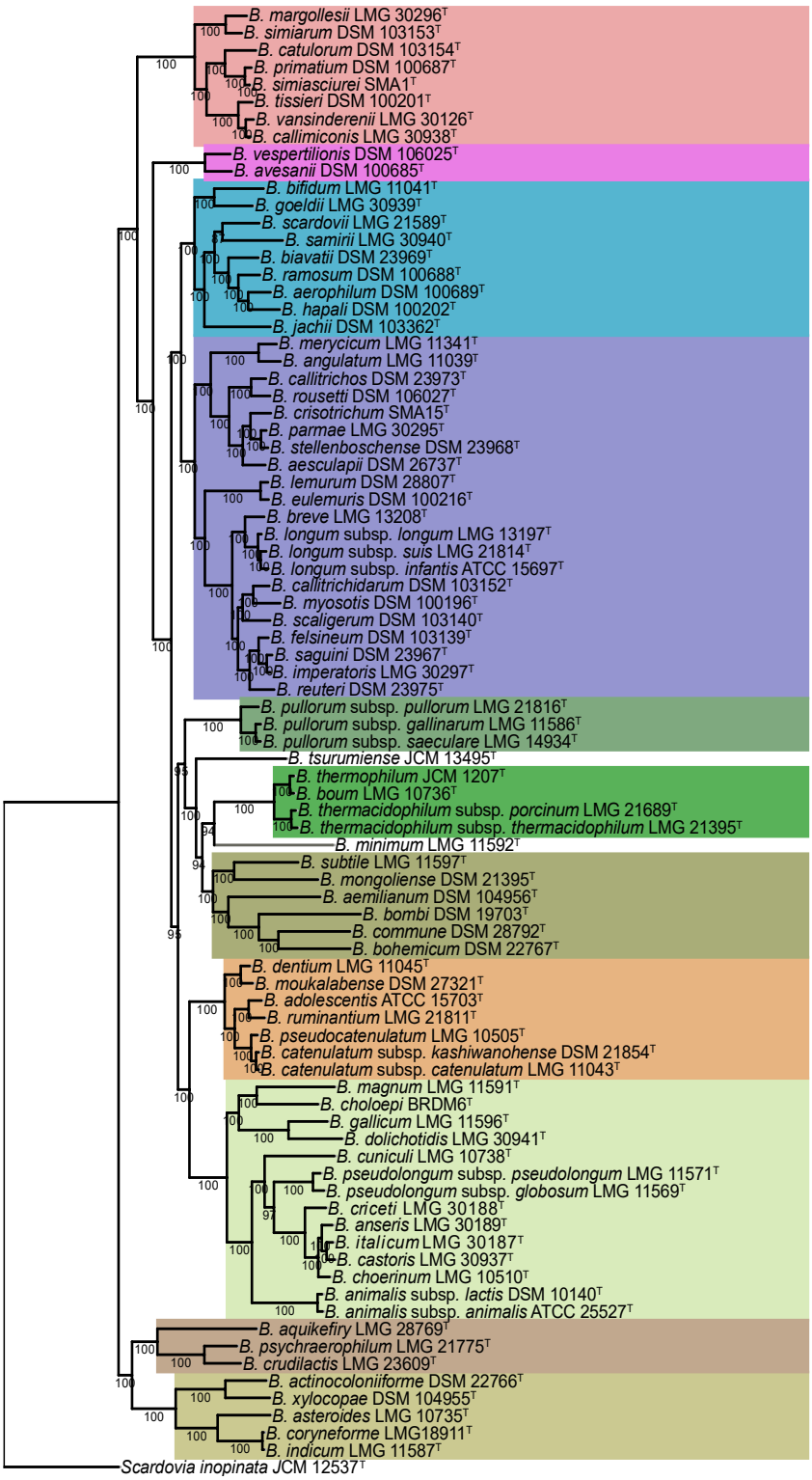


Figure 5.1. Phylogenetic tree based on concatenated amino acid sequences of 362 core genes of the 84 type strains. Bootstrap percentages of >70 are shown. Eleven phylogenetic groups are highlighted in different colors and the new group is the second rightmost (rose) [46].

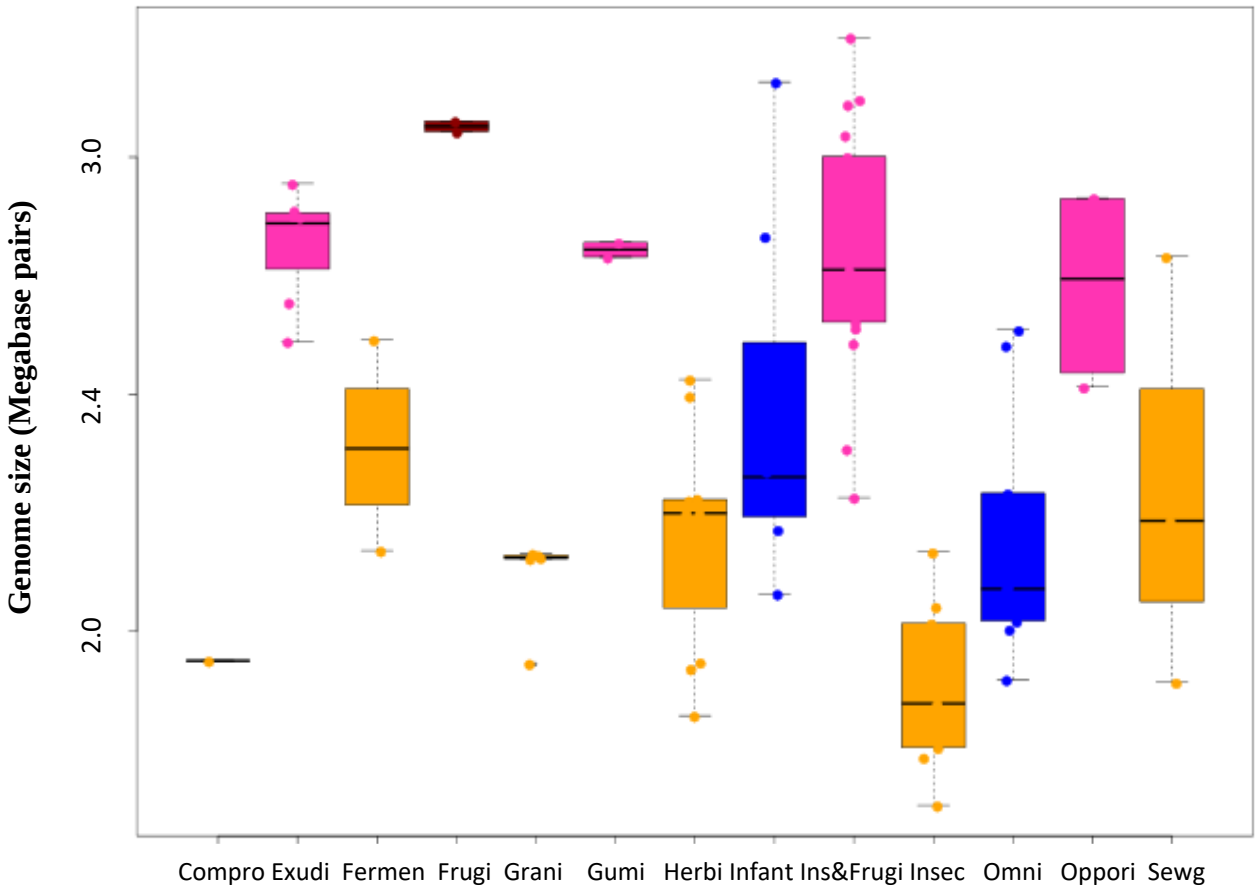


Figure 5.2. Genome sizes of the strains in each dietary group. The box plot indicates the mean and standard deviation. Compro: Commercial probiotic; Exudi: Exudativore; Fermen: Fermented food; Frugi: Frugivore; Grani: Granivore; Gumi: Gummivore; Herbi: Herbivore; Infant: Infant food; Ins&Frugi: Frugivore eating insects; Insec: Nectarivore, palynivore; Omni: Omnivore; Oppori: Opportunistic omnivore eating fruits, leaves and insects; Sewg: Sewage. Exudi, gumi, and grani eat insects too. The colors in the boxplot shows different host groups; Dark red: bats, Pink: monkey/apes, Blue: human/pigs, Yellow: other animals [46].

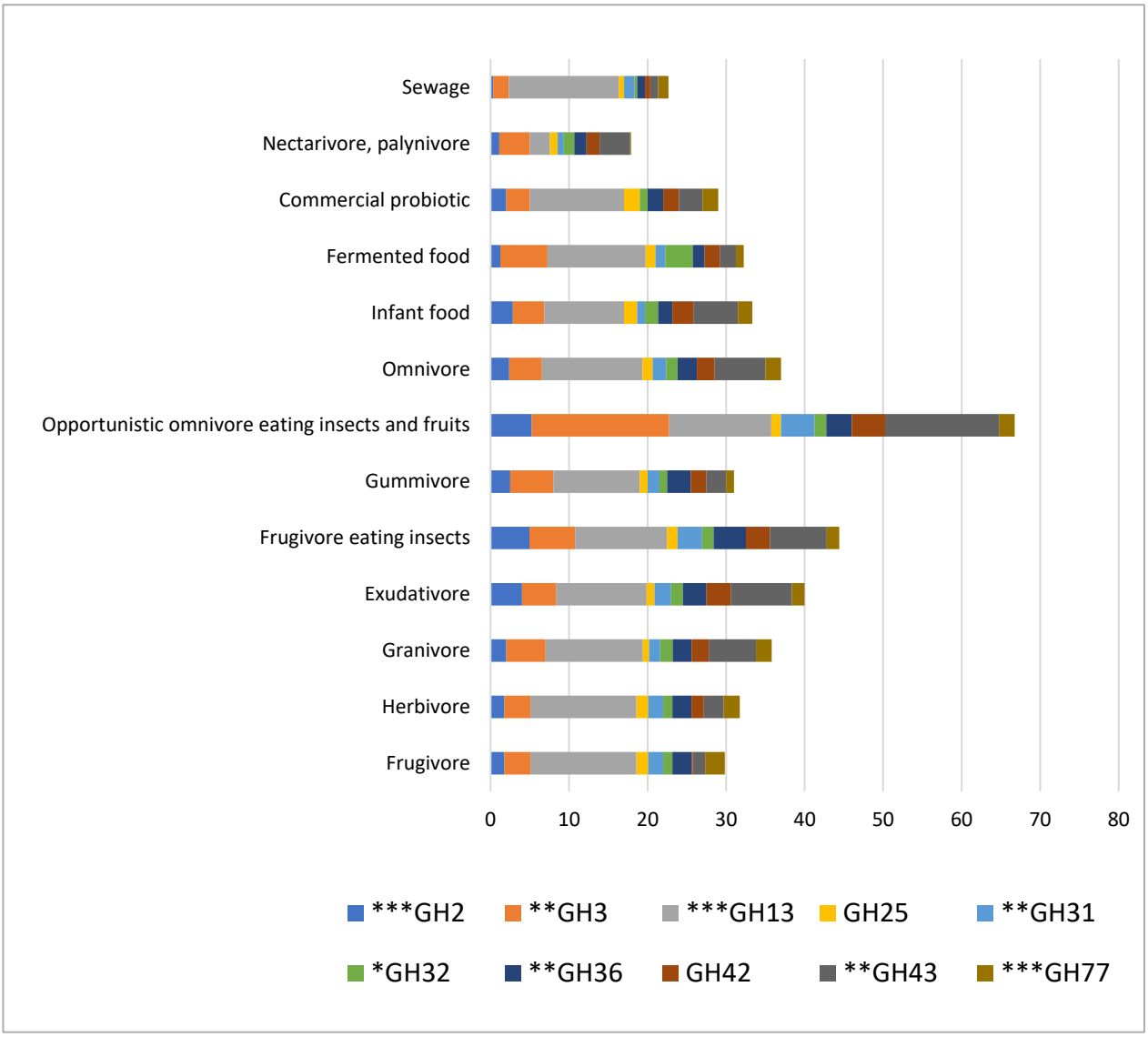


Figure 5.3. Distribution of abundances of active carbohydrate enzyme family's genes in the dietary groups. (**a**) Abundant glycoside hydrolase (GH) family genes. Major CAZyme families in >80% of the strains are shown. The significance by Kruskal-Wallis test is shown by asterisks. * p <0.05, ** p <0.01, *** p <0.001[46].

5.3.3. *Clustering of Bifidobacterium species based on GH families*

I next identified key GH families that delineate dietary difference of hosts. The clustering result of GH families became stable when 32 families that were present in >20% of all strains

were used (see Methods). The clustering created Group I-V in Figure 5.4, with the following characteristic families (Table 5.2).

1. Group I included strains with the largest number of GH genes. This group reflected species from opportunistic omnivore eating insects and fruits. The group had high numbers of GH43 and GH3 genes associated with degradation of complex plant polysaccharides like xylan, arabinan or arabinoxylan degradation. This suggested that these GH genes were adapted to the hosts of mixed diets (omnivore and frugivore).
2. Group II included strains with a high number of GH43 but low GH3. The group included 25 species and was further divided into three: Group II A, B and C. The subgroup II-C possessed low numbers of GH2, GH28, GH59 and GH115. The dietary pattern of the hosts varied: omnivore, herbivore, frugivore, insectivore and exudativore.
3. Group III included bee isolates and two infant isolates. This group possessed a very low number of GH13. This result was supported by previous studies where the GHs from the insects clustered separately [26]. GH13 enzymes are involved in degradation of starches and malto-oligosaccharides, and such sugars are usually scarce in diets of bees and infants.
4. Group IV included strains from hosts of insect and fruit diet. This group had the second highest gene counts for GHs after Group I, which suggested that the species from frugivorous hosts possessed more GH genes.
5. Group V included the largest number of strains. This group had the lowest GH gene counts, where many of the GH families were mostly absent (e.g. no GH28, GH38 and GH115). The group was further divided into two subgroups (Group V-A and V-B). Group V-B was strains from herbivorous hosts while Group V-A included strains from hosts of mixed dietary habits.

Table 5.2. CAZy family's characteristic to different dietary groups ($p < 0.05$) [46].

Dietary Groups	CAZy Families	Related activities
Opportunistic omnivore eating insects and fruits and	GH13	α-1,4-glucosidase,amylopullulanase,sucrose Phosphorylase,α-amylase

Table 5.2. (Continued)

Frugivore eating insects (Group I, Group II-B and Group IV)	GH3	β-glucosidase,β-hexosaminidase
	GH43	Endo-1,5-α-L-arabinosidase,α-L-arabinofuranosidase, Endo-1,4-β-xylanase,β-1,4-xylosidase
	GH26	Endo-1,4-β-mannosidase
	GH53	Endogalactanase
	GH31	α-xylosidase
	GH78	α-L-rhamnosidase
	CBM67	L-rhamnose binding activity
Frugivore eating insects (Group II-B and Group IV)	GH115	xylan α-1,2-glucuronidase,α-(4-O-methyl)-glucuronidase
	GH28	Galacturan1,4-α-galacturonidase,pectinesterase
Herbivore (Group V-B)	GH94	Cellobiose-phosphorylase
	GH36	α-galactosidase,raffinose synthase
Infant food (Group II-C)	GH33	Sialidase
	GH20	β-hexosaminidase
	GH29	α-L-fucosidase
	GH95	α-L-fucosidase
	GH112	Lacto-N-biosephosphorylase

Table 5.2. (Continued)

Nectarivore and Palynivore (Group III)	GH29	α-L-fucosidase
	GH95	α-L-fucosidase
	GH65	α,α-trehalase
	GH13*	α-1,4-glucosidase,amylopullulanase,sucrose Phosphorylase,α-amylase
	GT20	α,α-trehalose-phosphate synthase
	GT35*	glycogen or starch phosphorylase
	CBM48*	appended to GH13 modules
	CE10*	arylesterase

In Figure 5.4, the strains from insectivorous and frugivorous hosts were spread in separate clusters (Group IV and Group II). This discrepancy was attributed to the strains isolated from tamarins, whose diet is mainly insects and fruits but sometimes small amphibians. When the host diet was more complex (e.g. opportunistic omnivore, and frugivore and folivore), more diverse GH families and more genes were found. On the contrary, the strains from hosts with simple feeding habits (e.g. pure herbivore and nectarivore) possessed smaller number of families and genes. A good example was four subspecies of *B. longum*: subsp. *longum*, subsp. *suis*, subsp. *infantis*, and subsp. *suillum*. Of the three subspecies whose genomes were available, the former two belonged to Group II, while subsp. *infantis* belonged to Group III, due to different diets of their hosts. Hosts of the subsp. *longum* and subsp. *suis* are omnivores, while subsp. *infantis* is only seen in human infants. Infants generally consume simple diet, including breast milk and infant formulae, and thus storage of numerous GHs is not essential for the strain.

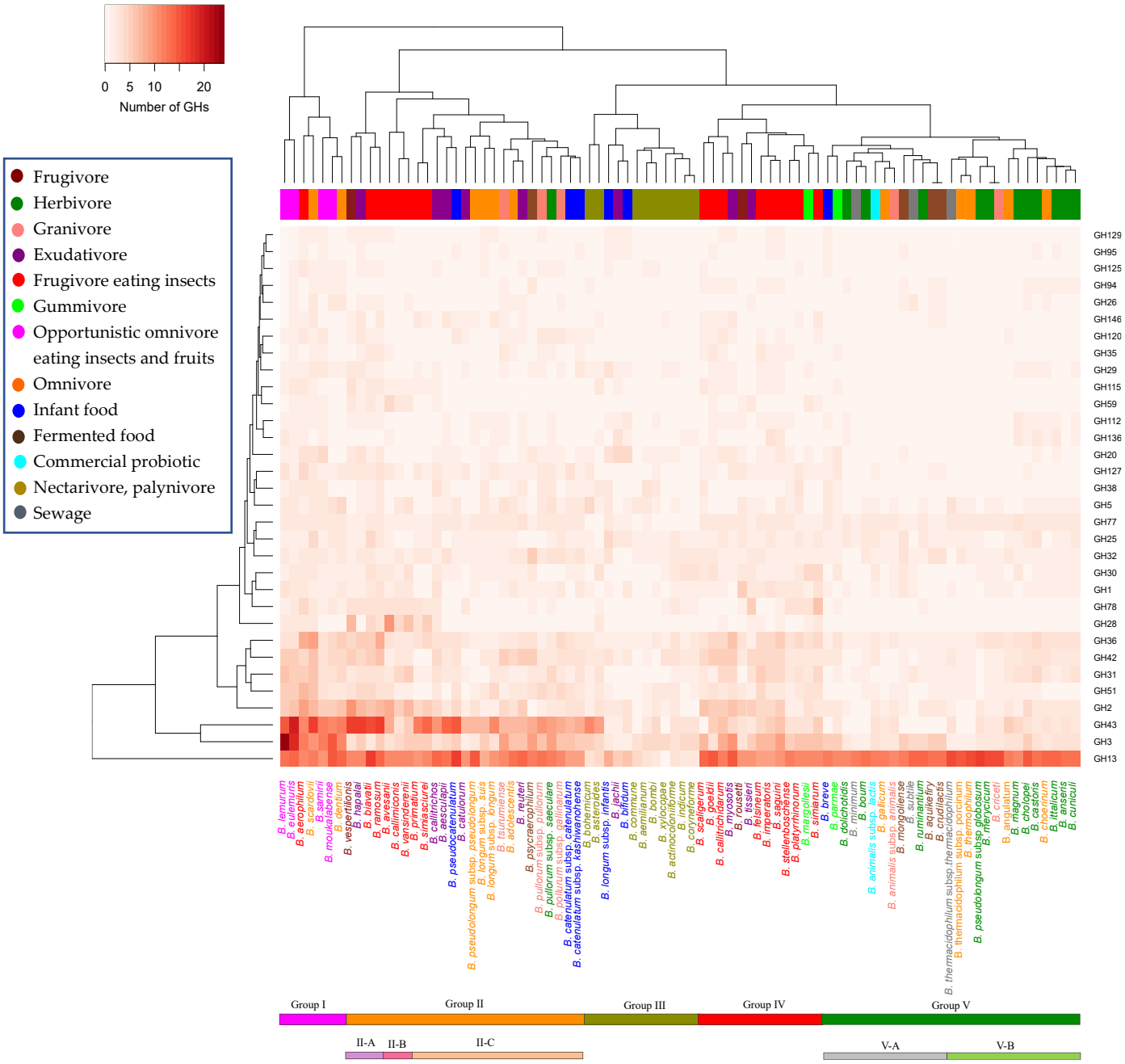
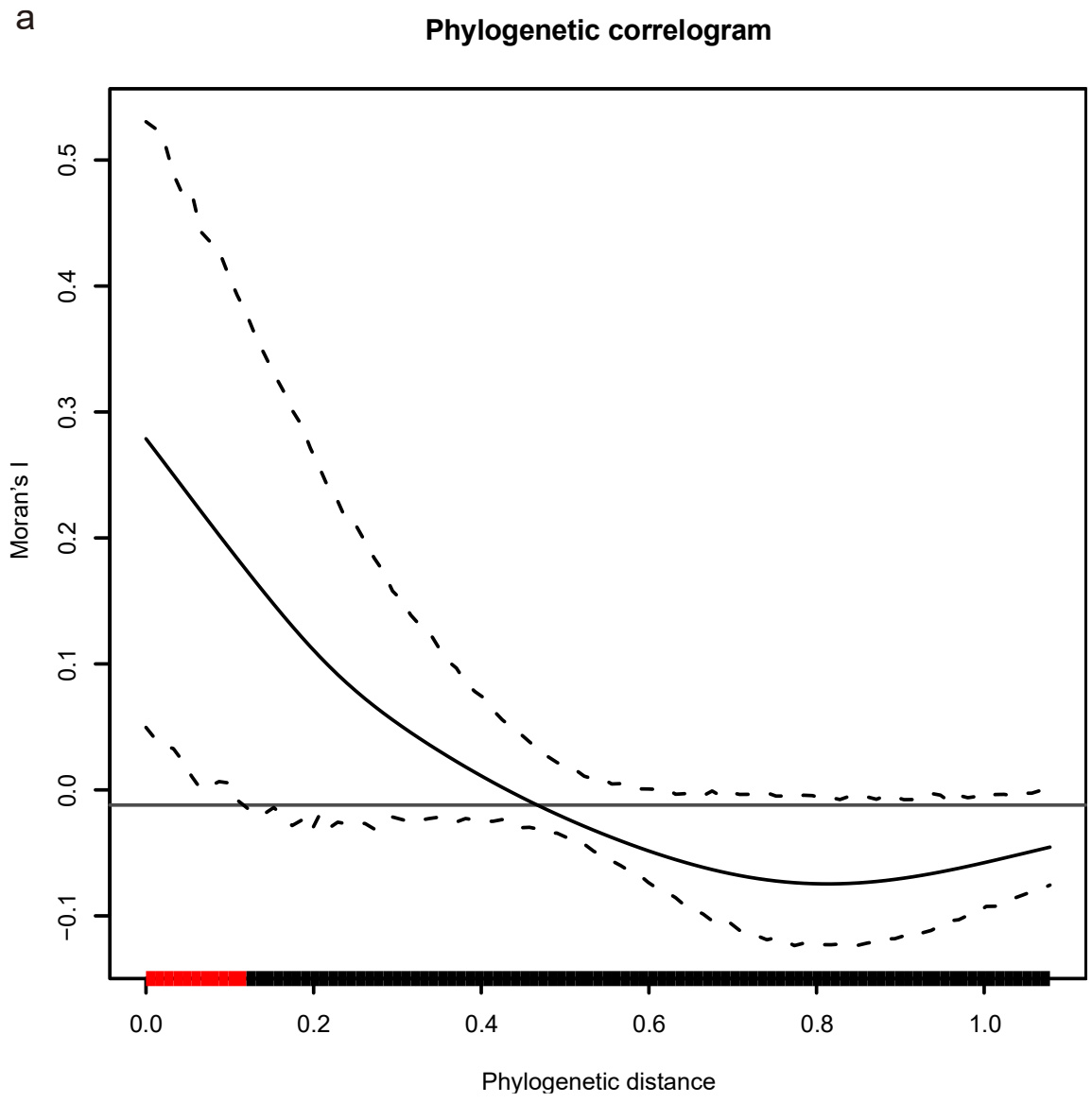


Figure 5.4. Clustering of bifidobacterial species based on GH family genes. The heatmap shows the gene number for the selected GH families (families present in 20% of the strains). Pink: Group I with the opportunistic omnivores; Orange: Group II with omnivore, herbivore or insectivore; Gold: Group III with nectarivore; Red: Group IV with insectivore and frugivore; Green: Group V with herbivore and mixed diet. Each strain is highlighted with the colour of the corresponding diet class [46].

To test whether the GH contents follow the dietary pattern rather than the phylogeny, I checked the phylogenetic signal for GH genes. The analysis showed weak phylogenetic signal with Bloomberg's K value closer to 0 (K = 0.448). Phylogenetic correlogram analysis detected nonsignificant autocorrelation above the phylogenetic distance of 0.1 (Figure 5.5a). I also performed the LIPA analysis to identify clades with a high phylogenetic signal. Only two clades (Clade 1: *Bifidobacterium eulemuris* and *Bifidobacterium lemurum*; Clade 2: *Bifidobacterium hapali*, *Bifidobacterium aerophilium*, *Bifidobacterium ramosum*, *Bifidobacterium biavatii*, *B. scardovii*, and *Bifidobacterium samirii*) were detected with significant positive autocorrelation (*p*-value < 0.01) (Figure 5.5b).



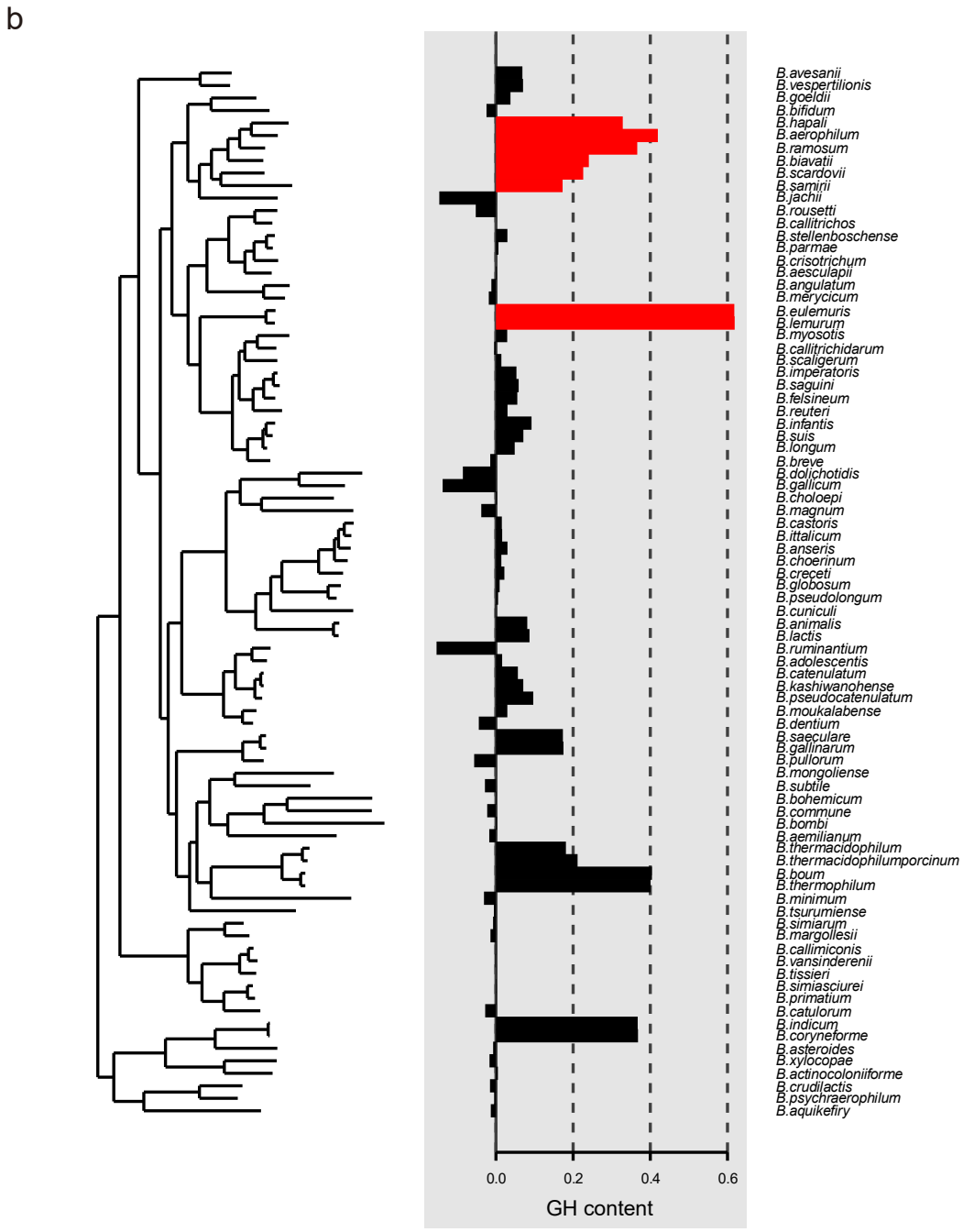


Figure 5.5. (a) Phylogenetic correlogram based on GH content. No significant autocorrelation was observed above the phylogenetic distance of 0.1. The dash lines represent the lower and upper confidence intervals and solid line represents the Moran's I index of autocorrelation. The colored horizontal bars at the bottom shows the significance of autocorrelation: red- significant positive autocorrelation, blue – significant negative autocorrelation and black – no autocorrelation. **(b)** Local Moran's index values for GH content for each type strain. The clades highlighted in red shows the presence of significant phylogenetic signal for p-value <0.01 [46].

5.3.4. *Comparison of Bifidobacterium species from multiple host animals*

Some species were isolated from multiple host animals of different dietary patterns. To investigate their GHs, I selected 66 strains in 11 different species isolated from different hosts (Supplementary Table 5.2). Their clustering resulted in 7 different groups, from Cluster (i) to Cluster (vii), among which five groups (Cluster (i), (iv), (v), (vi), and (i)) cleanly corresponded to the species' phylogeny (*Bifidobacterium moukalabense*, *B. breve*, *Bifidobacterium thermophilum*, *B. pseudolongum*, and *B. animalis*) (Figure 5.6).

The result suggested that strains within the same species shared similar GH families. Still, we could find characteristic GH families that coincided with host diet patterns. For example, *B. moukalabense* strains from gorilla, chimpanzee, and elephant possessed high numbers of GH families for plant carbohydrates (GH43, GH3, GH13, GH53, GH26 and GH78). *B. thermophilum* from pig, cow, and human lacked GH43 and GH2, and these families hydrolyze plant carbohydrates and milk carbohydrates, respectively (Table 5.3). *B. bifidum* strains were isolated from infants and calf and possessed high numbers of GH families for milk-origin carbohydrates (GH2, GH20, GH33, GH129 and GH84). Among the milk carbohydrates was GH33 for sialidase, whose abundance is statistically significant in *B. bifidum*, *B. longum* subsp. *infantis*, and *B. breve* only [162].

To further investigate the variation of GH genes within the same species, I selected 45 strains of *B. animalis* subsp. *lactis* from 15 different isolation sources (Supplementary Table 5.3). Many strains were isolated from humans probably because of extensive use of probiotic strains (re-isolation). The clustering based on GH genes within subsp. *lactis* showed a single large isogenic group with a small isolated group from dog, pig and food products (Supplementary Figure 5.4). This result supported that strains in the same species share similar GH patterns. The reason for the large deviation of some strains may be due to an application of unique strains as probiotics for animals. When all available *B. animalis* subsp. *lactis* strains were investigated for their GH genes, the 95% confidence interval for the number of GH genes in each family was never larger than 0.4. This indicated that the number of GH genes did not differ much within the same species and justified our approach of using type strains to grasp the overview of metabolic capabilities in *Bifidobacterium.*

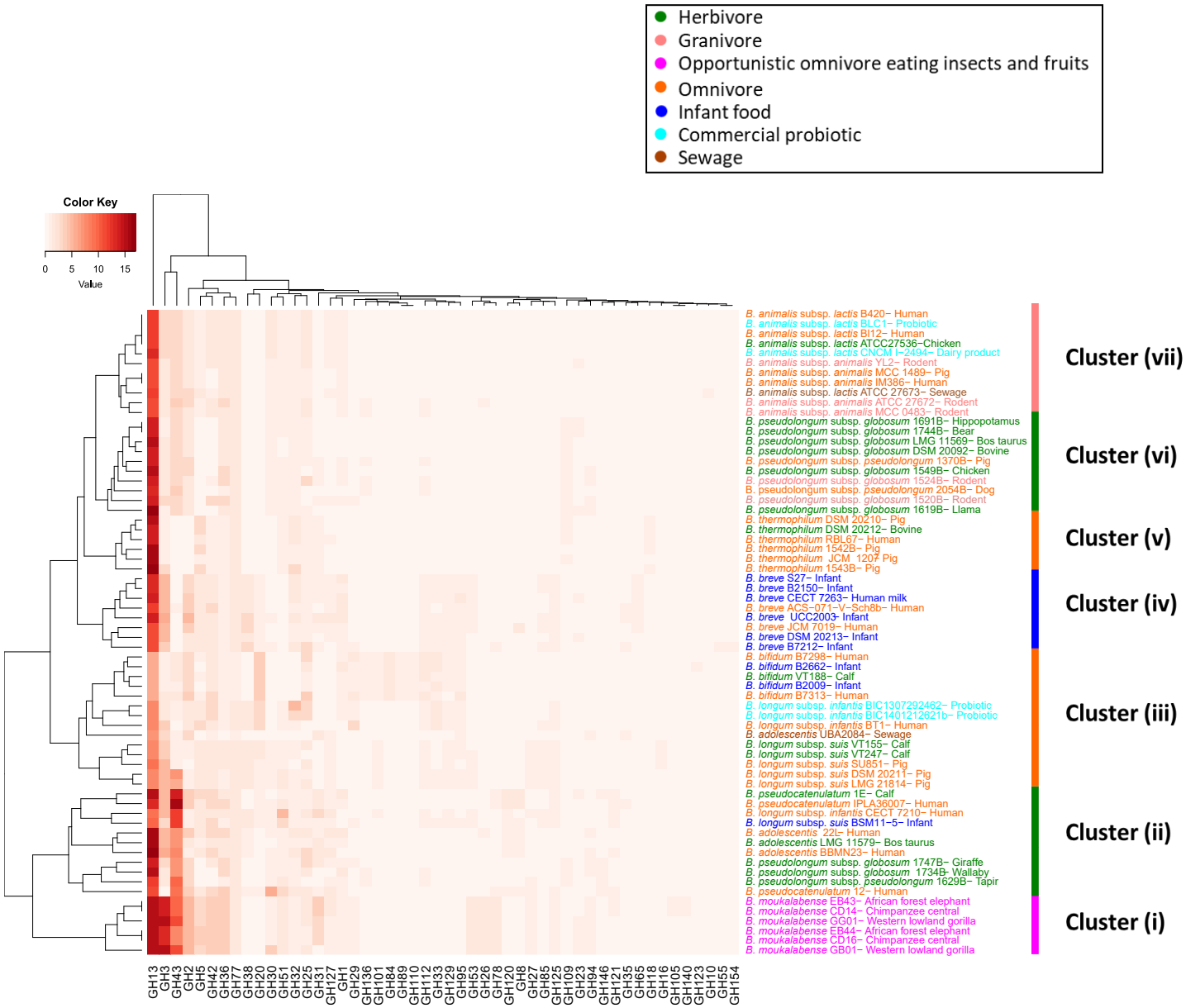


Figure 5.6. Clustering of 66 strains isolated from different sources based on their GHs. Heatmap displays the number of genes in GH families. Strains were colored according to their host dietary patterns as in the upper box. Strains were clustered in seven major groups: Cluster (i) Opportunistic omnivore; Cluster (ii) and Cluster (vi) Herbivore; Cluster (iii) and Cluster (v) Omnivore; Cluster (iv) Infant food; and Cluster (vii) Granivore and Insectivore [46].

Table 5.3. Characteristic GH families in the *Bifidobacterium* species with multiple host ($p < 0.05$) [46].

Family	Related subfamilies	Significantly high	Significantly low
GH1	β-glucosidase,β-galactosidase	*B. bifidum*	*B. longum* subsp. *suis*
GH2	β-galactosidase	all others	*B. thermophilum*
GH3	β-glucosidase,β-hexosaminidase, β-glucosideglucohydrolase	*B. thermophilum*, *B. bifidum*	*B. moukalabense*
GH5	β-mannosidase,β-glucosidase, β-exoglucanase	*B. moukalabense*	*B. pseudolongum* subsp. *globosum*
GH13	α-1,4-glucosidase, amylopullulanase, sucrose phosphorylase,α-amylase	*B. moukalabense*	*B. bifidum*
GH20	β-hexosaminidase	*B. bifidum*	all others
GH26	Endo-1,4-β-mannosidase	*B. moukalabense*	all others
GH27	α-galactosidase	*B. moukalabense*	all others
GH29	α-L-fucosidase	*B. bifidum*	*B. thermophilum*
GH30	β-D-xylosidase,endo-1,6-β-glucosidase, Glucosylceramidase	all others	*B. thermophilum*
GH31	α-xylosidase	*B. moukalabense*	all others
GH32	β-fructofuranosidase,sucrose-6-phosphatehydrolase	all others	*B. bifidum*
GH33	Sialidase	*B. bifidum*	*B. pseudolongum* subsp. *globosum*
GH36	α-galactosidase,raffinosesynthase	*B. moukalabense*	*B. thermophilum*
GH43	Endo-1,5-α-L-arabinosidase,α-L-arabinofuranosidase, Endo-1,4-β-xylanase,β-1,4-xylosidase	all others	*B. thermophilum*
GH51	α-L-arabinofuranosidase	*B. moukalabense*	*B. bifidum*
GH53	Endogalactanase	*B. moukalabense*	*B. pseudolongum* subsp. *globosum*
GH77	4-α-glucanotransferase	*B. bifidum*	all others
GH78	α-L-rhamnosidase	*B. moukalabense*	all others
GH84	α-L-rhamnosidase	*B. bifidum*	all others
GH85	Endo-β-N-acetylglucosaminidase D	*B. longum* subsp. *suis*	all others
GH89	α-N-acetylglucosaminidase,β-N-hexosaminidase	*B. bifidum*	all others
GH94	Cellobiose-phosphorylase	*B. moukalabense*	all others
GH95	α-L-fucosidase	*B. bifidum*	all others
GH101	endo-α-N-acetylgalactosaminidase	*B. bifidum*	all others
GH109	α-N-acetylgalactosaminidase	*B. pseudolongum* subsp. *globosum*	all others
GH110	Exo-α-galactosidase	*B. bifidum*	all others
GH112	Lacto-N-biosephosphorylase	*B. bifidum*	all others
GH120	β-xylosidase	*B. pseudocatenulatum*	all others
GH121	β-galactosidase	*B. pseudocatenulatum*	all others
GH127	β-L-arabinofuranosidase	*B. moukalabense*	all others

5.4. Conclusions

Genome-based features can deepen the understanding of the bacterial adaptation with host. I classified *Bifidobacterium* strains into five groups based on their GH genes, and the key GH families delineated the differences in host diet. The species from hosts having complex dietary habits possessed considerably more GH genes than those having simpler dietary patterns. Furthermore, a weak phylogenetic signal was confirmed for the distribution of GH genes.

In summary, bifidobacteria are adapted to their hosts' dietary habits, and their GH composition is associated with the diet composition. However, the GH composition within the same species did not match the host diet well. The shuffling speed of GH genes is therefore not faster than the speciation and host adaptation.

CHAPTER 6

General Discussion and Conclusion

In this study, four research projects aiming to investigate the interaction and adaptation of bifidobacteria to its diverse host range using a comparative genomic approach are reported. The first is a preliminary study focused to investigate the characteristics of genus *Bifidobacterium*, providing accurate genomic annotations and selecting a core genome. The second study inspects the host interaction and immunomodulatory role of bifidobacteria by investigating one of the important *B. bifidum* strain TMC3115. The last two studies focused on relationship of bifidobacteria with its host diet.

The primary findings of this study based on comparative analysis of genus *Bifidobacterium* with another probiotic genus *Lactobacillus* revealed the metabolic characteristics of genus *Bifidobacterium*. The protein families overrepresented in *Bifidobacterium* were found to be mostly involved in complex sugar metabolism host interaction, and stress responses. The analysis also showed more niche adjusted metabolic activities in *Bifidobacterium* such as broad adaptability for amino acids and polysaccharide metabolism.

Further the investigation of an important B. *bifidum* strain TMC3115 provided insight into the extracellular structures which might have their role in host interaction and immunomodulation. The study highlighted that there is variability among the genomes just not on species level but also on strain level in terms of host interaction.

The major finding in this work is that the bifidobacteria are adapted to their hosts' dietary habits, and their GH composition is associated with the diet composition. Here I investigated the relationship between bifidobacteria and their host diet using a comparative genomics approach. Since carbohydrates are the main class of nutrients for bifidobacterial growth, I examined the distribution of carbohydrate-active enzymes, in particular glycoside hydrolases (GHs) that metabolize unique oligosaccharides. When bifidobacterial species are classified by their distribution of GH genes, five groups arose according to their hosts' feeding behaviour. The distribution of GH genes was only weakly associated with the phylogeny of

the host animals or with genomic features such as genome size. Thus, the hosts' dietary pattern is the key determinant of the distribution and evolution of GH genes.

This study as a whole provides insight into bifidobacterial adaptation to its ecological niches. The reference library, the new statistical method for core genome selection and the findings obtained in this study can be further used to elucidate the genomic characteristics of this important genus.

References

1. Thursby E, Juge N. Introduction to the human gut microbiota. *Biochemical Journal.* 2017;474(11):1823-1836.
2. Salminen SJ, Gueimonde M, Isolauri E. Probiotics that modify disease risk. *The Journal of nutrition.* 2005;135(5):1294-1298.
3. Turroni F, Peano C, Pass DA, Foroni E, Severgnini M, Claesson MJ, et al. Diversity of bifidobacteria within the infant gut microbiota. *PLoS one.* 2012;7(5): e36957.
4. Favier CF, de Vos WM, Akkermans AD. Development of bacterial and bifidobacterial communities in feces of newborn babies. *Anaerobe.* 2003;9(5):219-229.
5. van Bergeijk DA, Terlouw BR, Medema MH, van Wezel GP. Ecology and genomics of Actinobacteria: new concepts for natural product discovery. *Nature Reviews Microbiology*. 2020;18(10):546-558.
6. Ventura M, Canchaya C, Tauch A, Chandra G, Fitzgerald GF, Chater KF, van Sinderen D. Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiology and molecular biology reviews*. 2007;71(3):495-548.
7. Lee JH, O'Sullivan DJ. Genomic insights into bifidobacterial. *Microbiology and Molecular Biology Reviews.* 2010;74(3):378-416.
8. Leahy SC, Higgins DG, Fitzgerald GF, van Sinderen D. Getting better with bifidobacteria. *Journal of applied microbiology.* 2005;98(6):1303-1315.
9. Alberoni D, Gaggìa F, Baffoni L, Modesto M, Biavati B, Gioia D. Bifidobacterium xylocopae sp. nov. and Bifidobacterium aemilianum sp. nov., from the carpenter bee (Xylocopa violacea) digestive tract. *Systematic and applied microbiology*. 2019;42(2):205-216.
10. Modesto M, Watanabe K, Arita M, Satti M, Oki K, Sciavilla P, et al. Bifidobacterium jacchi sp. nov., isolated from the faeces of a baby common marmoset (Callithrix jacchus). *International journal of systematic and evolutionary microbiology*. 2019;69(8):2477-2485.
11. Modesto M, Satti M, Watanabe K, Puglisi E, Morelli L, Huang CH, et al. Characterization of Bifidobacterium species in feaces of the Egyptian fruit bat: Description of B. vespertilionis sp. nov. and B. rousetti sp. Nov. *Systematic and applied microbiology*. 2019;42(6):126017.
12. Duranti S, Lugli GA, Napoli S, Anzalone R, Milani C, Mancabelli L, 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. *International journal of systematic and evolutionary microbiology*. 2019;69(5):1288-1298.

13. Lugli GA, Mangifesta M, Duranti S, Anzalone R, Milani C, Mancabelli L, et al. Phylogenetic classification of six novel species belonging to the genus Bifidobacterium comprising Bifidobacterium anseris sp. nov., Bifidobacterium criceti sp. nov., Bifidobacterium imperatoris sp. nov., Bifidobacterium italicum sp. nov., Bifidobacterium margollesii sp. nov. and Bifidobacterium parmae sp. nov. *Systematic and applied microbiology*. 2018;41(3):173-183.
14. Trovatelli LD, Crociani F, Pedinotti M, Scardovi V. Bifidobacterium pullorum sp. nov.: a new species isolated from chicken feces and a related group of bifidobacteria isolated from rabbit feces. *Archives of microbiology*. 1974;98(1):187-198.
15. Biavati B, Scardovi V, Moore WE. Electrophoretic patterns of proteins in the genus Bifidobacterium and proposal of four new species. *International Journal of Systematic and Evolutionary Microbiology*. 1982;32(3):358-373.
16. Lugli GA, Milani C, Duranti S, Alessandri G, Turroni F, Mancabelli L, et al. Isolation of novel gut bifidobacteria using a combination of metagenomic and cultivation approaches. *Genome biology*. 2019;20(1):96.
17. Holzapfel WH, Wood BJ, editors. Lactic acid bacteria: biodiversity and taxonomy. *John Wiley & Sons*. 2014;p.521.
18. Turroni F, Van Sinderen D, Ventura M. Genomics and ecological overview of the genus Bifidobacterium. *International journal of food microbiology*. 2011;149(1):37-44.
19. Holzapfel WH, Haberer P, Geisen R, Björkroth J, Schillinger U. Taxonomy and important features of probiotic microorganisms in food and nutrition. *The American journal of clinical nutrition*. 2001;73(2):365s-373s.
20. Peres CM, Peres C, Hernández-Mendoza A, Malcata FX. Review on fermented plant materials as carriers and sources of potentially probiotic lactic acid bacteria–With an emphasis on table olives. *Trends in Food Science & Technology.* 2012;26(1):31-42.
21. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight, R, et al. Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences*. 2011;108(Suppl.1):4578–4585.
22. Avershina E, Lundgard K, Sekelja M, Dotterud C, Storro O, Oien T, et al. Transition from infant- to adult-like gut microbiota. *Environmental microbiology*. 2016;18(7):2226–2236.

23. Turroni F, Peano C, Pass DA, Foroni E, Severgnini M, Claesson MJ, et al. Diversity of bifidobacteria within the infant gut microbiota. *PLoS one*. 2012;7(5):e36957.
24. Ishikawa E, Matsuki T, Kubota H, Makino H, Sakai T, Oishi K, et al. Ethnic diversity of gut microbiota: Species characterization of Bacteroides fragilis group and genus Bifidobacterium in healthy Belgian adults, and comparison with data from Japanese subjects. *Journal of bioscience and bioengineering*. 2013;116(2):265–270.
25. Odamaki T, Bottacini F, Kato K, Mitsuyama E, Yoshida K, Horigome A, et al. Genomic diversity and distribution of Bifidobacterium longum subsp. longum across the human lifespan. *Scientific reports*. 2018;8(1):1-12.
26. Turroni F Foroni, E Pizzetti, P Giubellini V, Ribbera A,Merusi P, et al. Exploring the diversity of the bifidobacterial population in the human intestinal tract. *Applied and environmental microbiology*. 2009;75(6):1534–1545.
27. Milani C, et al. Genomic encyclopedia of type strains of the genus *Bifidobacterium*. *Applied and environmental microbiology.* 2014;80(20): 6290–6302.
28. Pokusaeva K, Fitzgerald GF, Van Sinderen D. Carbohydrate metabolism in Bifidobacteria. *Genes & nutrition.* 2011;6(3):285–306.
29. Sánchez B, Urdaci MC, Margolles A. Extracellular proteins secreted by probiotic bacteria as mediators of effects that promote mucosa-bacteria interactions. *Microbiology*. 2010;156(11): 3232–3242.
30. Fukuda, S, Toh, H, Hase, K, Oshima, K, Nakanishi, Y, Yoshimura, K, et al. Bifidobacteria can protect from enteropathogenic infection through production of acetate. *Nature*. 2011;469(7331):543-547.
31. Correa NB, Peret Filho LA, Penna FJ, Lima FMS, Nicoli JR. A randomized formula controlled trial of *Bifidobacterium lactis* and *Streptococcus thermophilus* for prevention of antibiotic-associated diarrhea in infants. *Journal of clinical gastroenterology*. 2005;39(5):385–389.
32. Patole SK, Rao SC, Keil AD, Nathan EA, Doherty DA, Simmer KN. Benefits of Bifidobacterium breve M-16V supplementation in preterm neonates-a retrospective cohort study. *PloS one*. 2016;11(3):e0150775.

33. Gionchetti P, Rizzello F, Venturi A, Campieri M. Probiotics in infective diarrhoea and inflammatory bowel diseases. *Journal of gastroenterology and hepatology*. 2000;15(5):489-493

34. TANAKA R, TAKAYAMA H, MOROTOMI M, KUROSHIMA T, UEYAMA S, MATSUMOTO K, et al. Effects of administration of TOS and Bifidobacterium breve 4006 on the human fecal flora. *Bifidobacteria and microflora*, 1983;2(1):17-24.

35. Charnchai P, Jantama SS, Prasitpuriprecha C, Kanchanatawee S, Jantama K. Effects of the food manufacturing chain on the viability and functionality of *Bifidobacterium animalis* through simulated gastrointestinal conditions. *PLoS One.* 2016;11(6): e0157958.

36. Picard C, Fioramonti J, Francois A, Robinson T, Neant F, Matuchansky C. bifidobacteria as probiotic agents–physiological effects and clinical benefits. *Alimentary pharmacology & therapeutics*. 2005;22(6):495-512.

37. Pompei A, Cordisco L, Amaretti A, Zanoni S, Matteuzzi D, Rossi M. Folate production by bifidobacteria as a potential probiotic property. *Applied and environmental microbiology*. 2007;73(1):179-185.

38. Wong CB, Odamaki T, Xiao JZ. Insights into the reason of Human-Residential Bifidobacteria (HRB) being the natural inhabitants of the human gut and their potential health-promoting benefits. *FEMS microbiology reviews*. 2020;*44*(3):369-385.

39. Duranti S, Longhi G, Ventura M, van Sinderen D, Turroni F. Exploring the Ecology of Bifidobacteria and Their Genetic Adaptation to the Mammalian Gut. *Microorganisms*. 2021;9(1):8.

40. Schell, MA, Karmirantzou, M, Snel, B, Vilanova, D, Berger, B, Pessi, G, et al. The genome sequence of Bifidobacterium longum reflects its adaptation to the human gastrointestinal tract. *Proceedings of the National Academy of Sciences*. 2002;99(22): 14422-14427.

41. Satti M, Tanizawa Y, Endo A, Arita M. Comparative analysis of probiotic bacteria based on a new definition of core genome. *Journal of bioinformatics and computational biology*. 2018;16(03):1840012.

42. Philippe H, Douady CJ. Horizontal gene transfer and phylogenetics. *Current opinion in microbiology*. 2003;6(5):498–505.

43. Ventura M, Canchaya C, Casale AD, Dellaglio F, Neviani E, Fitzgerald GF, et al. Analysis of bifidobacterial evolution using a multilocus approach. *International journal of systematic and evolutionary microbiology*. 2006;56(12):2783–2792.
44. Lugli GA, Milani C, Duranti S, Mancabelli L, Mangifesta M, Turroni F, et al. Tracking the taxonomy of the genus Bifidobacterium based on a phylogenomic approach. *Applied and environmental microbiology*. 2018;84(4):e02249-17.
45. Lugli GA, Milani C, Turroni F, Duranti S, Ferrario C, Viappiani A, et al. Investigation of the evolutionary development of the genus Bifidobacterium by comparative genomics. *Applied and environmental microbiology*. 2014;80(20):6383-6394.
46. Satti M, Modesto M, Endo A, Kawashima T, Mattarelli P, Arita M. Host-Diet Effect on the Metabolism of *Bifidobacterium*. *Genes*. 2021;12(4):609.
47. Alessandri G, Ossiprandi MC, MacSharry J, van Sinderen D, Ventura M. Bifidobacterial dialogue with its human host and consequent modulation of the immune system. *Frontiers in immunology*. 2019;10:2348.
48. Ventura M, Turroni F, Motherway MOC, MacSharry J, van Sinderen D. Host–microbe interactions that facilitate gut colonization by commensal bifidobacteria. *Trends in microbiology*. 2012;*20*(10):467-476.
49. Foroni E, Serafini F, Amidani D, Turroni F, He F, Bottacini F, et al. Genetic analysis and morphological identification of pilus-like structures in members of the genus Bifidobacterium. *Microbial Cell Factories BioMed Central*. 2011;10(1):1-13.
50. Milani C, Mangifesta M, Mancabelli L, Lugli GA, Mancino W, Viappiani A, et al. The sortase-dependent fimbriome of the genus *Bifidobacterium*: extracellular structures with potential to modulate microbe-host dialogue. *Applied and environmental microbiology*. 2017;83(19).
51. Turroni F, Serafini F, Foroni E, Duranti S, Motherway MO, Taverniti V, et al. Role of sortase-dependent pili of Bifidobacterium bifidum PRL2010 in modulating bacterium-host interactions. *Proceedings of the National Academy of Sciences*. 2013;110(27):11151-11156.
52. Ventura M, Turroni F, Motherway MO, MacSharry J, van Sinderen D. Host-microbe interactions that facilitate gut colonization by commensal bifidobacteria. *Trends in microbiology*. 2012;20(10):467-476.
53. O'Connell Motherway M, Houston A, O'Callaghan G, Reunanen J, O'Brien F, O'Driscoll T, et al. A Bifidobacterial pilus-associated protein promotes colonic epithelial proliferation. *Molecular microbiology*. 2019;111(1):287-301.

54. Turroni F, Ventura M, Butto LF, Duranti S, O'Toole, PW Motherway MO, et al. Molecular dialogue between the human gut microbiota and the host: A Lactobacillus and Bifidobacterium perspective. *Cellular and Molecular Life Sciences*. 2014;71(2):183–203.
55. Horn N, Wegmann, U Dertli, E Mulholland,F, Collins SR, Waldron KW, et al. Spontaneous mutation reveals influence of exopolysaccharide on Lactobacillus johnsonii surface characteristics. *PloS one*. 2014;*8*(3):e59957.
56. Fanning S, Hall LJ, van Sinderen D. Bifidobacterium breve UCC2003 surface exopolysaccharide production is a beneficial trait mediating commensal-host interaction through immune modulation and pathogen protection. *Gut microbes*. 2012;3(5):420-425.
57. Ferrario C, Milani C, Mancabelli L, Lugli GA, Duranti S, Mangifesta M, et al. Modulation of the eps-ome transcription of bifidobacteria through simulation of human intestinal environment. *FEMS microbiology ecology*. 2016;92(4).
58. Fanning S, Hall LJ, Cronin M, Zomer A, MacSharry J, Goulding D, et al. Bifidobacterial surface-exopolysaccharide facilitates commensal-host interaction through immune modulation and pathogen protection. *Proceedings of the National Academy of Sciences*. 2012;109(6):2108-2113.
59. Turroni F, Foroni E, Motherway MOC, Bottacini F, Giubellini V, Zomer A, et al. Characterization of the serpin-encoding gene of Bifidobacterium breve 210B. *Applied and environmental microbiology*. 2010;76(10):3206-3219.
60. Sela DA, Chapman J, Adeuya A, Kim JH, Chen F, Whitehead TR, et al. The genome sequence of Bifidobacterium longum subsp. infantis reveals adaptations for milk utilization within the infant microbiome. *Proceedings of the National Academy of Sciences*. 2008;105(48):18964-18969.
61. Turroni F, Bottacini F, Foroni E, Mulder I, Kim JH, Zomer A, et al. Genome analysis of Bifidobacterium bifidum PRL2010 reveals metabolic pathways for host-derived glycan foraging. *Proceedings of the National Academy of Sciences*. 2010;107(45): 19514-19519.
62. Duranti S, Milani C, Lugli GA, Mancabelli L, Turroni F, Ferrario C, et al. Evaluation of genetic diversity among strains of the human gut commensal Bifidobacterium adolescentis. *Scientific reports*. 2016;6(1):1-10.
63. Schell MA, Karmirantzou M, Snel B, Vilanova D, Berger B, Pessi G, et al. The genome sequence of Bifidobacterium longum reflects its adaptation to the human gastrointestinal tract. *Proceedings of the National Academy of Sciences*. 2002;99(22): 14422-14427.
64. Maze A, O'Connell-Motherway, M, Fitzgerald GF, Deutscher J, van Sinderen D. Identification and characterization of a fructose phosphotransferase system

in *Bifidobacterium breve* UCC2003. *Applied and environmental microbiology*. 2007;73(2):545–553.

65. Parche S, Amon J, Jankovic I, Rezzonico E, Beleut M, Barutçu H, et al. Sugar transport systems of Bifidobacterium longum NCC2705. *Journal of molecular microbiology and biotechnology*. 2007;12(1-2):9-19.
66. Turroni F, Strati F, Foroni E, Serafini F, Duranti S, van Sinderen D, et al. Analysis of predicted carbohydrate transport systems encoded by Bifidobacterium bifidum PRL2010. *Applied and environmental microbiology*. 2012;78(14):5002-5012.
67. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B, et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic acids research*. 2009;37(suppl_1):D233-D238.
68. Lombard V, Bernard T, Rancurel C, Brumer H, Coutinho PM, Henrissat B. A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochemical Journal*. 2010;432(3):437- 444.
69. Sheridan PO, Martin JC, Lawley TD, Browne HP, Harris HMB, Bernalier-Donadille A, Duncan SH, et al. Polysaccharide utilization loci and nutritional specialization in a dominant group of butyrate-producing human colonic Firmicutes. *Microbial Genomics*. 2016;2(2):e000043.
70. Milani C, Lugli GA, Duranti S, Turroni F, Bottacini F, Mangifesta M, et al. Genomic encyclopedia of type strains of the genus Bifidobacterium. *Applied and environmental microbiology*. 2014;80(20):6290-6302.
71. Milani C, Lugli GA, Duranti S, Turroni F, Mancabelli L, Ferrario C, et al. Bifidobacteria exhibit social behavior through carbohydrate resource sharing in the gut. *Scientific reports*. 2015;5(1):1-14.
72. El Kaoutari A, Armougom F, Gordon JI, Raoult D, Henrissat B. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nature Reviews Microbiology*. 2013;11(7):497-504.
73. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*. 2005;102(39):13950-13955.
74. Contreras-Moreira B, Vinuesa P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and environmental microbiology*. 2013;79(24): 7696-7701.

75. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research.* 2003;13(9):2178-2189.
76. Tanizawa Y, Fujisawa T, Kaminuma E, Nakamura Y, Arita M. DFAST and DAGA: web-based integrated genome annotation tools and resources. *Bioscience of microbiota, food and health.* 2016;35(4):173-184.
77. Sun Z, Zhang W, Guo C, Yang X, Liu W, Wu Y, et al. Comparative genomic analysis of 45 type strains of the genus *Bifidobacterium*: a snapshot of its genetic diversity and evolution. *PLoS One.* 2015;10(2):e0117912.
78. Park BH, Karpinets TV, Syed MH, Leuze MR, Uberbacher EC. CAZymes Analysis Toolkit (CAT): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. *Glycobiology.* 2010;20(12): 1574-1584.
79. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic acids research.* 2012;40(W1):W445-W451.
80. Salazar N, Prieto A, Leal JA, Mayo B, Bada-Gancedo JC, de los Reyes-Gavilán CG, et al. Production of exopolysaccharides by *Lactobacillus* and *Bifidobacterium* strains of human origin and metabolic activity of the producing bacteria in milk, *Journal of dairy science*. 2009;92(9):4158-4168.
81. Sun Z, He X, Brancaccio VF, Yuan J, Riedel CU. Bifidobacteria exhibit LuxS-dependent autoinducer 2 activity and biofilm formation. *PLoS One*. 2014;9(2): e88260.
82. Górska S, Dylus E, Rudawska A, Brzozowska E, Srutkova D, Schwarzer M, et al. Immunoreactive proteins of *Bifidobacterium longum* ssp. *longum* CCM 7952 and *Bifidobacterium longum* ssp. *longum* CCDM 372 identified by gnotobiotic mono-colonized mice sera, immune rabbit sera and non-immune human sera, *Frontiers in microbiology.* 2016;7:1537.
83. Fushinobu S. Unique sugar metabolic pathways of bifidobacterial. *Bioscience, biotechnology, and biochemistry*. 2011;74(12):2374-2384.
84. Slováková L, Dusková D, Marounek M. Fermentation of pectin and glucose, and activity of pectin-degrading enzymes in the rabbit caecal bacterium *Bifidobacterium pseudolongum*. *Letters in applied microbiology*. 2002;25(2):126-130.
85. Turroni F, Foroni E, Pizzetti P, Giubellini V, Ribbera A, Merusi P, et al. Exploring the diversity of the bifidobacterial population in the human intestinal tract. *Applied and environmental microbiology.* 2009;75(6):1534-1545.

86. Turroni F, Duranti S, Bottacini F, Guglielmetti S, Van Sinderen D, Ventura M. Bifidobacterium bifidum as an example of a specialized human gut commensal. *Frontiers in microbiology*. 2014;5:437.
87. López P, González-Rodríguez I, Gueimonde M, Margolles A, Suárez A. Immune response to Bifidobacterium bifidum strains support Treg/Th17 plasticity. *PLOS ONE*. 2011;6(9):e24776.
88. Turroni F, Taverniti V, Ruas-Madiedo P, Duranti S, Guglielmetti S, Lugli GA, et al. Bifidobacterium bifidum PRL2010 modulates the host innate immune response. *Applied and Environmental Microbiology*. 2014;80(2):730-740.
89. Furrie E, Macfarlane S, Kennedy A, Cummings JH, Walsh SV, O'neil DA, et al. Synbiotic therapy (Bifidobacterium longum/Synergy 1) initiates resolution of inflammation in patients with active ulcerative colitis: a randomised controlled pilot trial. *Gut*. 2005;54(2):242-249.
90. Ohno H, Tsunemine S, Isa Y, Shimakawa M, Yamamura H. Oral administration of Bifidobacterium bifidum G9-1 suppresses total and antigen specific immunoglobulin E production in mice. *Biological and Pharmaceutical Bulletin*. 2005;28(8):1462-1466.
91. Medina M, Izquierdo E, Ennahar S, Sanz Y. Differential immunomodulatory properties of Bifidobacterium logum strains: relevance to probiotic selection and clinical applications. *Clinical & Experimental Immunology*. 2007;150(3):531-538.
92. Sanchez B, Bressollier P, Urdaci MC. Exported proteins in probiotic bacteria: adhesion to intestinal surfaces, host immunomodulation and molecular cross-talking with the host. *FEMS Immunology & Medical Microbiology*. 2008;54(1):1-17.
93. Lebeer S, Claes I, Tytgat HL, Verhoeven TL, Marien E, von Ossowski I, et al. Functional analysis of Lactobacillus rhamnosus GG pili in relation to adhesion and immunomodulatory interactions with intestinal epithelial cells. *Applied and environmental microbiology*. 2012;78(1):185-193.
94. Bottacini F, Medini D, Pavesi A, Turroni F, Foroni E, Riley D, et al. Comparative genomics of the genus Bifidobacterium. *Microbiology*. (2010);*156*(11), 3243-3254.
95. Morita H, He F, Fuse T, Ouwehand AC, Hashimoto H, Hosoda M, et al. Adhesion of lactic acid bacteria to Caco-2 cells and their effect on cytokine secretion. *Microbiology and immunology*. 2002;46(4):293-297.
96. Harata G, He F, Takahashi K, Hosono A, Kawase M, Kubota A, et al. Bifidobacterium suppresses IgE-mediated degranulation of rat basophilic leukemia (RBL-2H3) cells, *Microbiology and immunology*. 2010;54(1):54-57.

97. Cheng R, Guo J, Pu F, Wan C, Shi L, Li H, et al. Loading ceftriaxone, vancomycin, and Bifidobacteria bifidum TMC3115 to neonatal mice could differently and consequently affect intestinal microbiota and immunity in adulthood. *Scientific reports.* 2019;9(1):1-15.
98. Krumsiek J, Arnold R, Rattei T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*. 2007;23(8):1026-1028.
99. Tada I, Tanizawa Y, Arita M. Visualization of consensus genome structure without using a reference genome. *BMC genomics*. 2017;18(2):1-9.
100. Noureen M, Tada I, Kawashima T, Arita M. Rearrangement analysis of multiple bacterial genomes. *BMC Bioinform.* 2019;20(23):1–10.
101. Okonechnikov K, Golosova O, Fursov M. Ugene Team. Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics*. 2012;28(8):1166–1167.
102. Luo H, Zhang CT, Gao F. (2014). Ori-Finder 2, an integrated tool to predict replication origins in the archaeal genomes. *Frontiers in microbiology*. 2014;5:482.
103. Hendrickson H, Lawrence JG. Mutational bias suggests that replication termination occurs near the dif site, not at Ter sites. *Molecular microbiology*. 2007;64(1):42-56.
104. Bertelli C, Laird M.R, Williams K.P, Simon Fraser University Research Computing Group, Lau BY, Hoad, et al. IslandViewer 4: Expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res.* 2017;45(W1):W30–W35.
105. Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. *Journal of molecular biology*. 2004;340(4):783-795.
106. Bendtsen JD, Kiemer L, Fausbøll A, Brunak S. Non-classical protein secretion in bacteria. *BMC microbiology*. 2005;5(1):58.
107. Krogh A, Larsson B, Von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes1. *Journal of molecular biology*. 2001;305(3):567-580.
108. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, etal. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics.* 2010;26(13):1608-1615.
109. Zhou M, Boekhorst J, Francke C, Siezen RJ. LocateP: genome-scale subcellular-location predictor for bacterial proteins. *BMC bioinformatics.* 2008;9(1):173.
110. Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A. Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Science*. 2003;12(8):1652-1662.

111. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, et al. The Pfam protein families database. *Nucleic acids research*. 2004;32(suppl_1): D138-D141, 2004.
112. Milani C, Mangifesta M, Mancabelli L, Lugli GA, Mancino W, Viappiani A, et al. The sortase-dependent fimbriome of the genus Bifidobacterium: extracellular structures with potential to modulate microbe-host dialogue. *Applied and environmental microbiology*. 2017:83(19).
113. Treangen TJ, Abraham AL, Touchon M, Rocha E P. Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS microbiology reviews*. 2009;33(3):539-571.
114. Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS microbiology reviews*. 2014;38(5):865-891.
115. Savic D J, Nguyen SV, McCullor K, McShan WM. Biological impact of a large-scale genomic inversion that grossly disrupts the relative positions of the origin and terminus loci of the Streptococcus pyogenes chromosome. *Journal of bacteriology*. 2019;201(17).
116. Guinane CM, Kent, RM, Norberg S, Hill C, Fitzgerald GF, Stanton C, Ross RP. Host specific diversity in Lactobacillus johnsonii as evidenced by a major chromosomal inversion and phage resistance mechanisms. *PLoS One*. 2011;6(4), e18740.
117. Delgado S, Sánchez B, Margolles A, Ruas-Madiedo P, Ruiz L. Molecules produced by probiotics and intestinal microorganisms with immunomodulatory activity. *Nutrients*. 2020;12(2):391.
118. Sanchez B, Urdaci MC, Margolles A. Extracellular proteins secreted by probiotic bacteria as mediators of effects that promote mucosa–bacteria interactions. *Microbiology*. 2010;156(11):3232-3242.
119. Ventura M, Turroni F, Lugli GA, van Sinderen D. Bifidobacteria and humans: our special friends, from ecological to genomics perspectives. *Journal of the science of food and agriculture*. 2014;94(2):163-168.
120. Wang G, Xia Y, Song X, Ai L. Common non-classically secreted bacterial proteins with experimental evidence. *Current microbiology*. 2016;72(1):102-111.
121. Guglielmetti S, Zanoni I, Balzaretti S, Miriani M, Taverniti V, De Noni I, et al. Murein lytic enzyme TgaA of Bifidobacterium bifidum MIMBb75 modulates dendritic cell maturation through its cysteine-and histidine-dependent amidohydrolase/peptidase (CHAP) amidase domain. *Applied and environmental microbiology*. 2014;80(17):5170-5177.

122. Bateman A, Holden MT, Yeats C. The G5 domain: a potential N-acetylglucosamine recognition domain involved in biofilm formation. *Bioinformatics*. 2004;21(8):1301-1303.
123. Boekhorst J, de Been MW, Kleerebezem M, Siezen RJ. Genome-wide detection and analysis of cell wall-bound proteins with LPxTG-like sorting motifs. *Journal of bacteriology*. 2005;187(14):4928-4934.
124. Nishiyama K, Yamamoto Y, Sugiyama M, Takaki T, Urashima T, Fukiya S, et al. Bifidobacterium bifidum Extracellular Sialidase Enhances Adhesion to the Mucosal Surface and Supports Carbohydrate Assimilation. *MBio*. 2017;8(5): e00928-17.
125. Turroni F, Bottacini F, Foroni E, Mulder I, Kim JH, Zomer A, et al. Genome analysis of Bifidobacterium bifidum PRL2010 reveals metabolic pathways for host-derived glycan foraging. *Proceedings of the National Academy of Sciences*. 2010;107(45):19514-19519.
126. Sánchez B, González-Tejedo C, Ruas-Madiedo P, Urdaci MC, Margolles A. Lactobacillus plantarum extracellular chitin-binding protein and its role in the interaction between chitin, Caco-2 cells, and mucin. *Applied and environmental microbiology* .2011;77(3):1123-1126.
127. Nakayama H, Kurokawa K, Lee BL. Lipoproteins in bacteria: structures and biosynthetic pathways. *The FEBS journal*. 2012;279(23):4247-4268.
128. Górska S, Dylus E, Rudawska A, Brzozowska E, Srutkova D, Schwarzer M, et al. Immunoreactive proteins of Bifidobacterium longum ssp. longum CCM 7952 and Bifidobacterium longum ssp. longum CCDM 372 Identified by gnotobiotic mono-colonized mice sera, immune rabbit sera and non-immune human sera. *Frontiers in microbiology*. 2016;7:1537.
129. Bottacini F, Motherway MOC, Kuczynski J, O'Connell KJ, Serafini F, Duranti S, et al. Comparative genomics of the Bifidobacterium breve taxon. *BMC genomics*. 2014;15(1):1-19.
130. Daveran-Mingot ML, Campo N, Ritzenthaler P, Le Bourgeois P.A Natural Large Chromosomal Inversion in Lactococcus lactis Is Mediated by Homologous Recombination between Two Insertion Sequences. *Journal of Bacteriology*. 1998;180(18):4834-4842.
131. Teeling EC, Vernes SC, Dávalos LM, Ray DA, Gilbert MT, Myers E, et al. Bat biology, genomes, and the Bat1K Project: To generate Chromosome-Level genomes for all living bat species. Annual review of animal biosciences. 2018;6**:**23-46.
132. Dietrich M, Markotter W. Studying the microbiota of bats: Accuracy of direct and indirect samplings. *Ecology and Evolution*. 2019;9(4):1730-1735.

133. Muegge BD, Kuczynski J, Knights D, Clemente JC, González A, Fontana L, et al. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science*. 2011;332(6032):970-974.
134. Kwiecinski GG, Griffiths TA. Rousettus egyptiacus. *Mammalian Species*. 1999;(611):1-9.
135. Hood WR, Oftedal OT, Kunz TH. Is tissue maturation necessary for flight? Changes in body composition during postnatal development in the big brown bat. *Journal of Comparative Physiology*. 2011;181(3):423-435.
136. Canani RB, Passariello A, Buccigrossi V, Terrin G, Guarino A. The nutritional modulation of the evolving intestine. *Journal of clinical gastroenterology*. 2008; Suppl 42:197-200.
137. Di Gioia D, Aloisio I, Mazzola G, Biavati B. Bifidobacteria: their impact on gut microbiota composition and their applications as probiotics in infants. *Applied microbiology and biotechnology*. 2014; 98(2):563-577.
138. O'Callaghan A, van Sinderen D. Bifidobacteria and their role as members of the human gut microbiota. *Frontiers in microbiology*. 2016;7: 925.
139. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of molecular biology*. 2016;428(4):726-731.
140. Saier Jr MH, Reddy VS, Tsu BV, Ahmed MS, Li C, Moreno-Hagelsieb G. The transporter classification database (TCDB): recent advances. *Nucleic acids research*. 2015;44(D1): D372-D379.
141. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30(4):772-780.
142. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;25(15):1972-1973.
143. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30(9):1312-1313.
144. Khoroshkin MS, Leyn SA, Van Sinderen D, Rodionov DA. Transcriptional regulation of carbohydrate utilization pathways in the Bifidobacterium genus. *Frontiers in microbiology*. 2016; 7:120.
145. Rodríguez-Díaz J, Monedero V, Yebra MJ. Utilization of natural fucosylated oligosaccharides by three novel α-l-fucosidases from a probiotic Lactobacillus casei strain. *Applied and environmental microbiology*. 2011;77(2):703- 705.

146.Korine C, Arad Z. Changes in milk composition of the Egyptian fruit bat, Rousettus aegyptiacus (Pteropodidae), during lactation. *Journal of mammalogy*. 1999;80(1):53-59.

147.Jenness RO, Studier EH. Lactation and milk. Special publications the museum texas tech university. 1976; 10:1-218.

148.Senda A, Kobayashi R, Fukuda K, Saito T, Hood WR, Kunz TH, Oftedal OT, Urashima T. Chemical characterization of milk oligosaccharides of the island flying fox (Pteropus hypomelanus) (Chiroptera: Pteropodidae). *Animal science journal*. 2011;82(6):782-786.

149.Kitaoka M. Bifidobacterial enzymes involved in the metabolism of human milk oligosaccharides. *Advances in nutrition*. 2012;3 Suppl 3:422-429.

150.Kelly WJ, Cookson AL, Altermann E, Lambie SC, Perry R, Teh KH, et al. Genomic analysis of three Bifidobacterium species isolated from the calf gastrointestinal tract. *Scientific reports*. 2016; 6:30768.

151.Garrido D, Dallas DC, Mills DA. Consumption of human milk glycoconjugates by infant-associated bifidobacteria: mechanisms and implications. *Microbiology*. 2013;159(4):649.

152.Yoshida E, Sakurama H, Kiyohara M, Nakajima M, Kitaoka M, Ashida H, et al. Bifidobacterium longum subsp. infantis uses two different β-galactosidases for selectively degrading type-1 and type-2 human milk oligosaccharides. *Glycobiology*. 2011;22(3):361-368.

153.O'Connell Motherway M, Fitzgerald GF, van Sinderen D. Metabolism of a plant derived galactose-containing polysaccharide by Bifidobacterium breve UCC2003. *Microbial biotechnology*. 2011;4(3):403-416.

154.Rodriguez CI, Martiny JB. Evolutionary relationships among bifidobacteria and their hosts and environments. *BMC genomics*. 2020;21(1):1-2.

155.Bottacini F, Milani C, Turroni F, Sánchez B, Foroni E, Duranti S, Serafini F, Viappiani A, Strati F, Ferrarini A, Delledonne M. *Bifidobacterium asteroides* PRL2011 genome analysis reveals clues for colonization of the insect gut. *PLoS One*. 2012;7(9):e44229.

156.Tanizawa Y, Fujisawa T, Nakamura Y. DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. *Bioinformatics*. 2018;34(6):1037-1039.

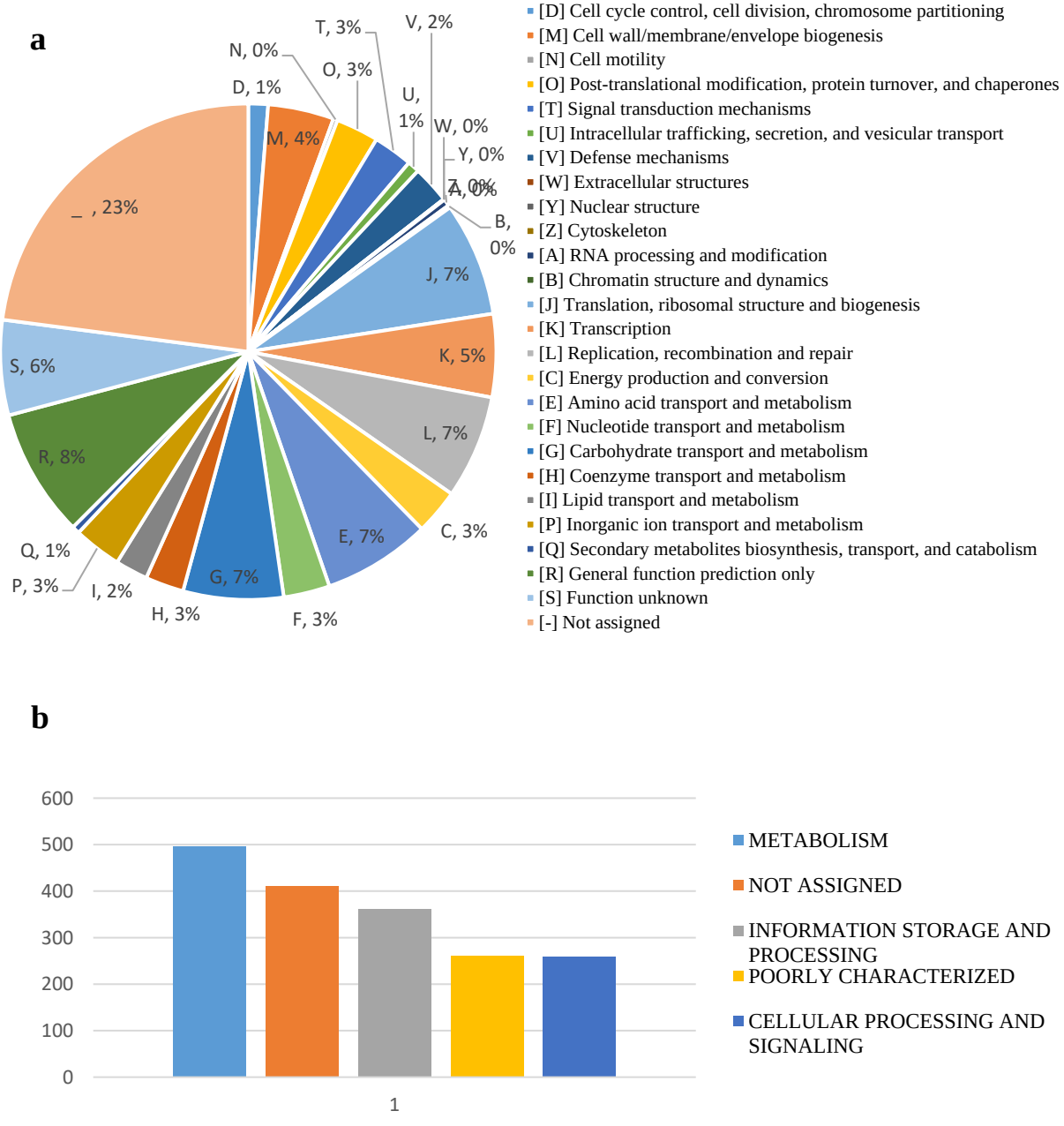
157.Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res*. 2018;46(W1):W95-W101.

158.Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res*. 2016;44(W1):W242-W245.

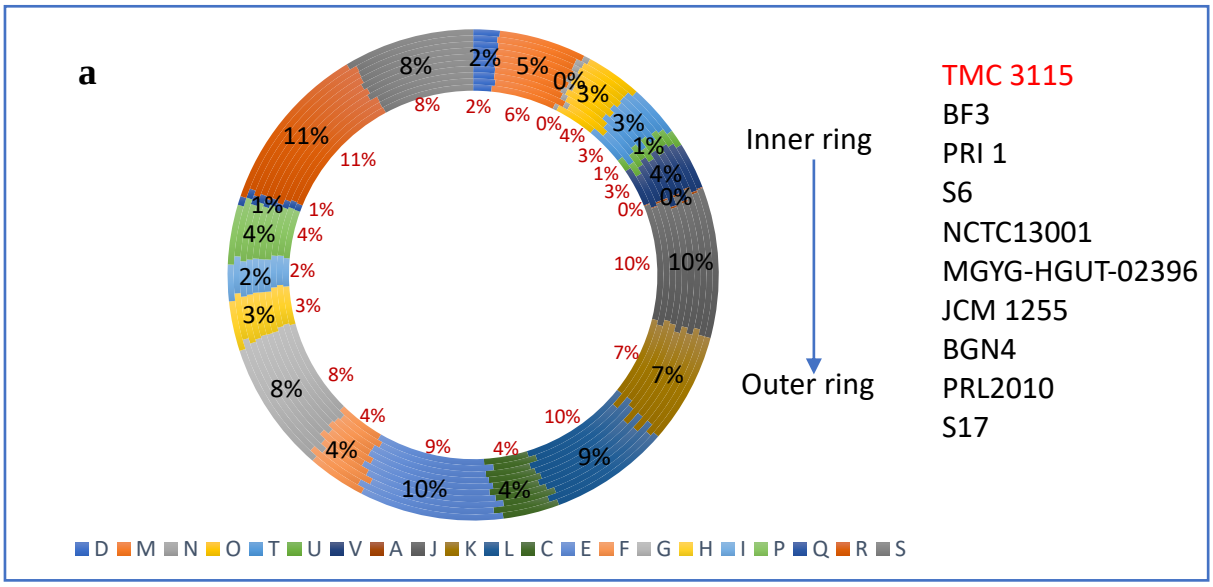
159.Keck F, Rimet F, Bouchez A, Franc A. phylosignal: an R package to measure, test, and explore the phylogenetic signal. *Ecol Evol.* 2016;6(9):2774-2780.
160.Blomberg SP, Garland Jr T, Ives AR. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*. 2003;57(4):717-745.
161.Stevens JR, Hallinan EV, Hauser MD. The ecology and evolution of patience in two New World monkeys. *Biology letters*. 2005;1(2):223-226.
162.Ward RE, Niñonuevo, M, Mills DA, L Lebrilla CB, German JB. In vitro fermentability of human milk oligosaccharides by several strains of bifidobacteria. *Mol Nutr Food Res*. 2007, 51(11):1398-1405.

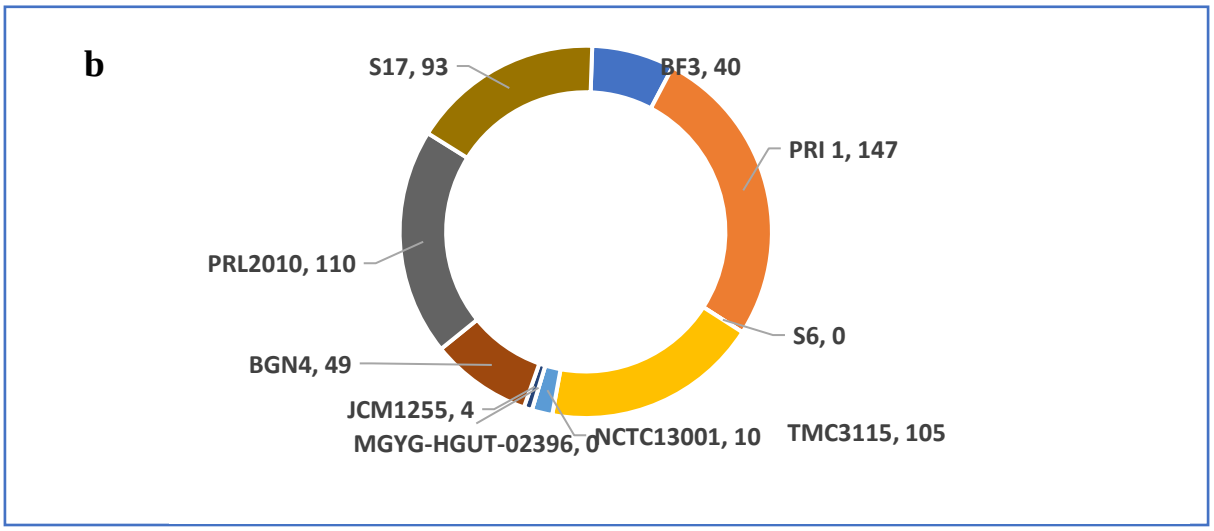
APPENDICES

APPENDIX 1. SUPPLEMENTARY MATERIAL FOR CHAPTER 3



Supplementary Figure 3.1. Distribution of Cluster of Orthologues (COG) functional categories in TMC3115 genome. **(a)** The COG subcategories distribution. **(b)** Top four COG categories distribution.





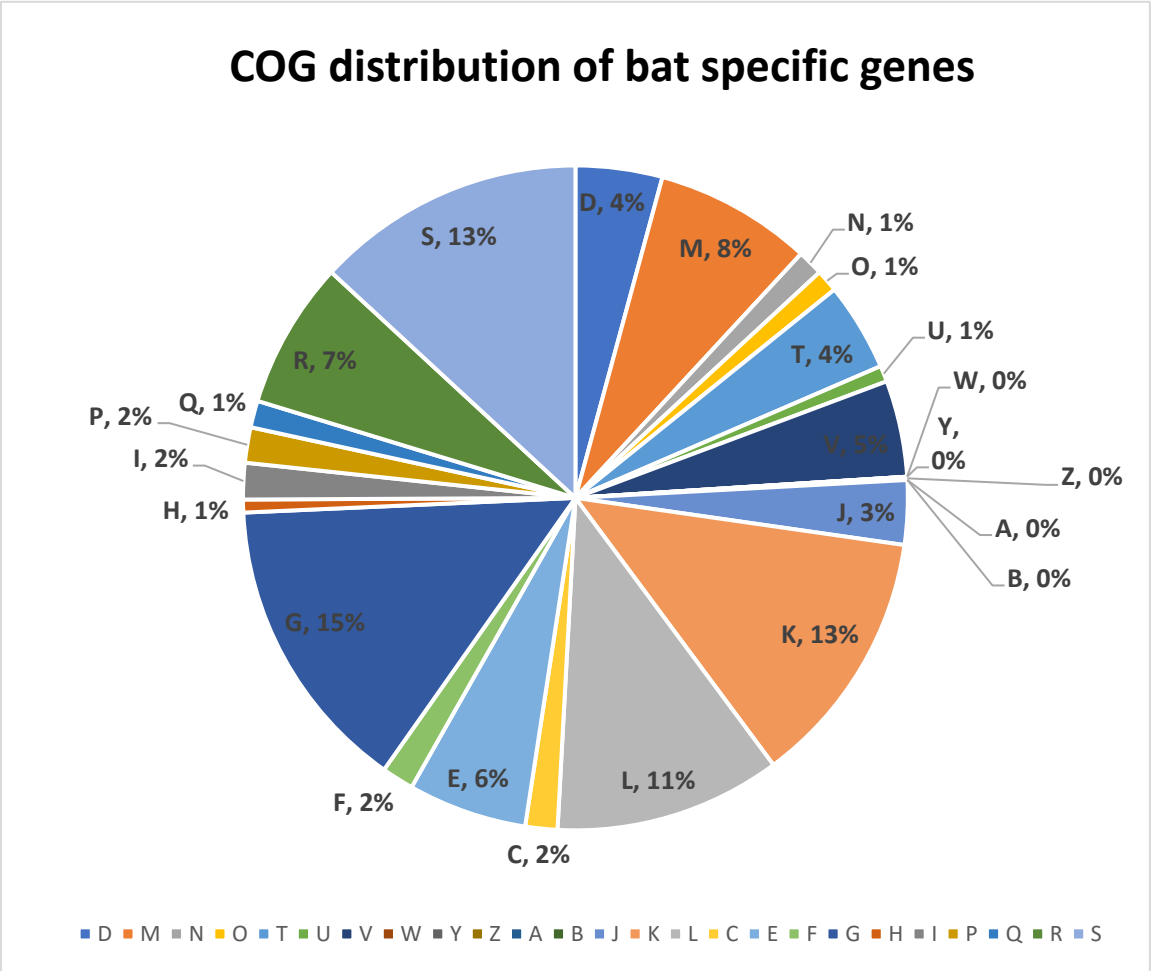
Supplementary Figure 3.2. Comparative genomics of *B. bifidum* TMC3115. **(a)** Distribution of COG categories among the strains. The numbers highlighted in black shows the average percentage of genes for each category while the number in red shows the percentage for the TMC3115 strain. COG classification: [D] Cell cycle control, cell division, chromosome partitioning; [M] Cell wall/membrane/envelope biogenesis;[N] Cell motility;[O] Post-translational modification, protein turnover, and chaperones;[T] Signal transduction mechanisms;[U] Intracellular trafficking, secretion, and vesicular transport;[V] Defense mechanisms;[A] RNA processing and modification; ;[J] Translation, ribosomal structure and biogenesis;[K] Transcription;[L] Replication, recombination and repair;[C] Energy production and conversion;[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;[P] Inorganic ion transport and metabolism;[Q] Secondary metabolites biosynthesis, transport, and catabolism;[R] General function prediction only;[S] Function unknown. **(b)** The number of unique genes present in each strain.

Supplementary Table 3.1. Sortase dependent pili clusters in *B. bifidum* strains. The strains are grouped in four groups based on number of pili and their pilin motifs.

STRAIN	No of Pili	MAJOR PILINS									Groups
		fimA			fimA			fimP			
		CWSS	Pilin Motif	E box	CWSS	Pilin Motif	E box	CWSS	Pilin Motif	E box	
PRL2010	3	LPGTG	GNATLTVSTK KGALPTVVKK	YTLTETEAPAGY	LPLTG	NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	G1
NCIMB 41171	3	LPGTG	KGALPTVVKK GNATLTVSTK GKTLLTVTMK	YTLTETEAPAGY	LPLTG	NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
BGN4	3	LPGTG	KGALPTVVKK NNNTLTVAMK	YTLTETEAPAGY	LPLTG	KGYQFTVSDK GTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
MJR8628B	3	LPGTG	KGALPTVVKK DNTLLTVAMK	YTLTETEAPAGY	LPLTG	NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VGTAATVTFK GGAAATVYAK	YTVTETAVADGY	
A8	3	LPGTG	KGNLPTVDKK NNNTLTVAMK	YTLTETEAPAGY	LPLTG	NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
324B	3	LPGTG	KGNLPTVDKK NNNTLTVAMK	YTLTETEAPAGY	LPLTG	NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
BF3	3	LPGTG	KGNLPTVDKK NNNTLTVAMK	YTLTETEAPAGY	LPLTG	NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
Bbif1887B	3	LPGTG	KGNLPTVDKK NNNTLTVAMK	YTLTETEAPAGY	LPLTG	NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
LMG 11582	3	LPGTG	KGDLPTVDKK NNNTLTVAMK	YTLTETEAPAGY	LPLTG	NGYQFTVSDK DTLKVTVDNK IGAGVTVGVK VGKTVTVEYK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
LMG 13195	3	LPGTG	KGDLPTVDKK NNNTLTVAMK	YTLTETEAPAGY	LPLTG	NGYQFTVSDK DTLKVTVDNK IGAGVTVGVK VGKTVTVEYK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
Calf96	3	LPGTG	KGDLPTVDKK DNTLLTVAMK	YTLTETEAPAGY	LPLTG	NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
S6	3	LPGTG	KGNLPTVDKK NNNTLTVAMK	YTLTETKAPAGY	LPLTG	NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
HGUT02396	3	LPGTG	KGNLPTVDKK NNNTLTVAMK	YTLTETKAPAGY	LPLTG	NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	

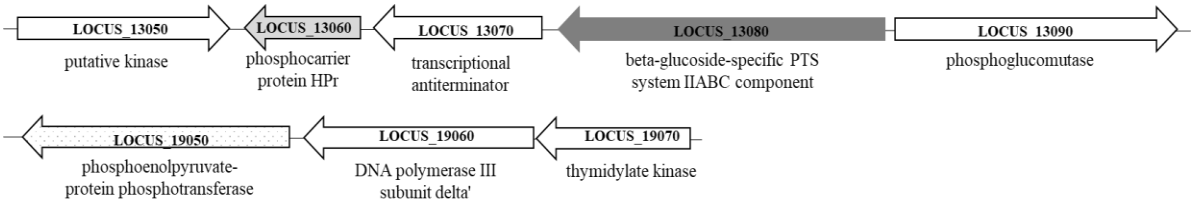
STRAIN	No of Pili	MAJOR PILINS									Groups
		fimA			fimA			fimP			
		CWSS	Pilin Motif	E box	CWSS	Pilin Motif	E box	CWSS	Pilin Motif	E box	
S17	3	LPGTG	KGDLPTVDKK	YTLTETEAPAGY	LPLTG	IGAGVTVGVK NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	G1
ATCC 29521	3	LPGTG	KGDLPTVDKK	YTLTETEAPAGY	LPLTG	NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
LMG 11041	3	LPGTG	KGDLPTVDKK	YTLTETEAPAGY	LPLTG	NGYQFTVSDK DTLKVTVDNK IGAGVTVGVK VGKTVTVEYK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
DSM 20456	3	LPGTG	KGDLPTVDKK	YTLTETEAPAGY	LPLTG	NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
JCM 1255	3	LPGTG	KGDLPTVDKK	YTLTETEAPAGY	LPLTG	NGYQFTVSDK DTLKVTVDNK IGAGVTVGVK VGKTVTVEYK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
NCTC13001	3	LPGTG	KGDLPTVDKK	YTLTETKAPAGY	LPLTG	NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
TMC3115	3	LPGTG	KGALPTVVKK NNNTLTVAMK	YTLTETEAPAGY	LPLTG	IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	G2
JCM 1254	3	LPGTG	NNNTLTVAMK	YTLTETEAPAGY	LKYTG	NGYQFTVSDK DTLKVTVDNK	YTIEEIAAPNGY	LPKTG	GGAAATVYAK	YTVTETAVADGY	
	3	LPGTG	KGDLPTVDKK NNNTLTVAMK	YTLTETEAPAGY	LKYTG	NGYQFTVSDK DTLKVTVDNK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
156B	2	LPGTG	KGALPTVVKK NNNTLTVAMK	YTLTETEAPAGY				LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	G3
ICIS-310	2	LPGTG	KGDLPTVDKK NNNTLTVAMK	YTLTETEAPAGY				LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
2789STDY560	2	LPGTG	KGNLPTVDKK NNNTLTVAMK	YTLTETEAPAGY				LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
791	2	LPGTG	KGNLPTVDKK NNNTLTVAMK	YTLTETEAPAGY				LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
BI-14	2	LPGTG	KGNLPTVDKK NNNTLTVAMK	YTLTETEAPAGY				LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
IPLA 20015	2	LPGTG	KGDLPTVDKK GNATLTVSTK	YTLTETEAPAGY				LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
85B	2	LPGTG	GNATLTVSTK KGDLPTVDKK	YTLTETEAPAGY				LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
IPLA 20017	2	LPGTG	KGDLPTVDKK GKTLLTVAMK	YTLTETEAPAGY				LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
LMG 11583	2	LPGTG	DNATLTVSTK KGALPTVVKK	YTLTETEAPAGY				LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
G1971	2	LPGTG	KGDLPTVDKK	YTLTETEAPAGY				LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
62-13	2				LPLTG	NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	G4
ASM157686v1	2				LPLTG	NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
CAG234	2				LPLTG	NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
PRI1	2				LPLTG	NGYQFTVSDK DTLKVTVDNK VGKNVTVEYK IGAGVTVGVK	YTIEEIAAPNGY	LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	
ASM157689v1	1							LPKTG	VDTAATVTFK GGAAATVYAK	YTVTETAVADGY	

APPENDIX 2. SUPPLEMENTARY MATERIAL FOR CHAPTER 4

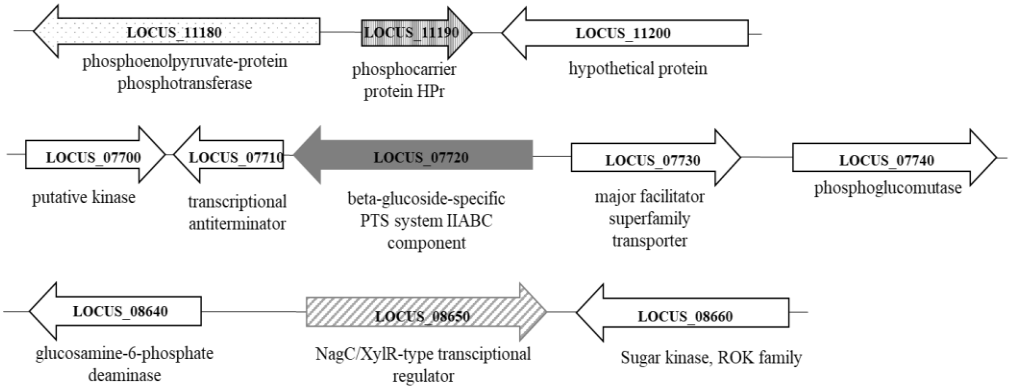


Supplementary Figure 4.1. COG distribution of bat specific genes. COG classification: [D] Cell cycle control, cell division, chromosome partitioning; [M] Cell wall/membrane/envelope biogenesis;[N] Cell motility;[O] Post-translational modification, protein turnover, and chaperones;[T] Signal transduction mechanisms;[U] Intracellular trafficking, secretion, and vesicular transport;[V] Defense mechanisms;[A] RNA processing and modification;[J] Translation, ribosomal structure and biogenesis;[K] Transcription;[L] Replication, recombination and repair;[C] Energy production and conversion;[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;[P] Inorganic ion transport and metabolism;[Q] Secondary metabolites biosynthesis, transport, and catabolism;[R] General function prediction only;[S] Function unknown.

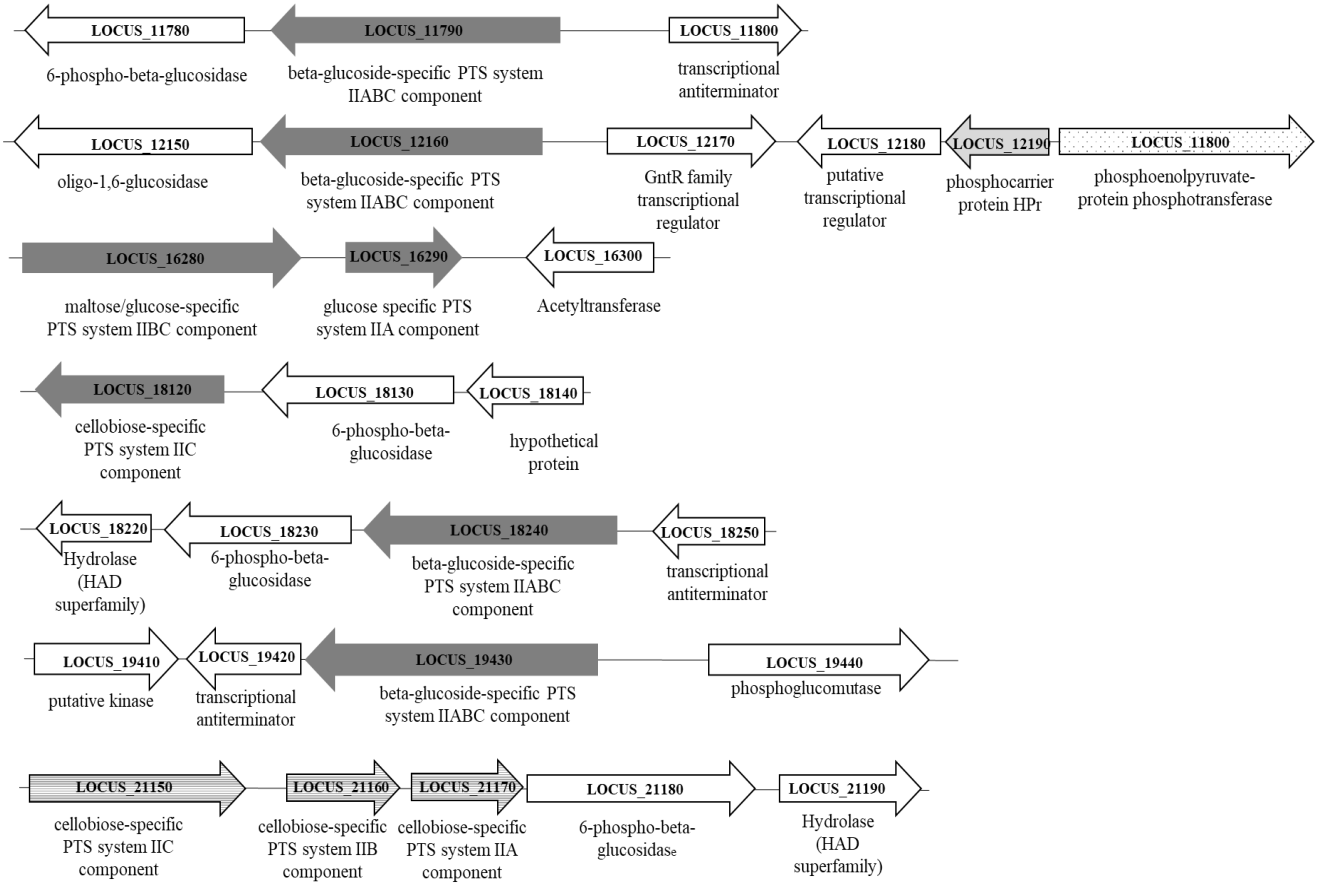
Cluster I- RST 7 & RST 11

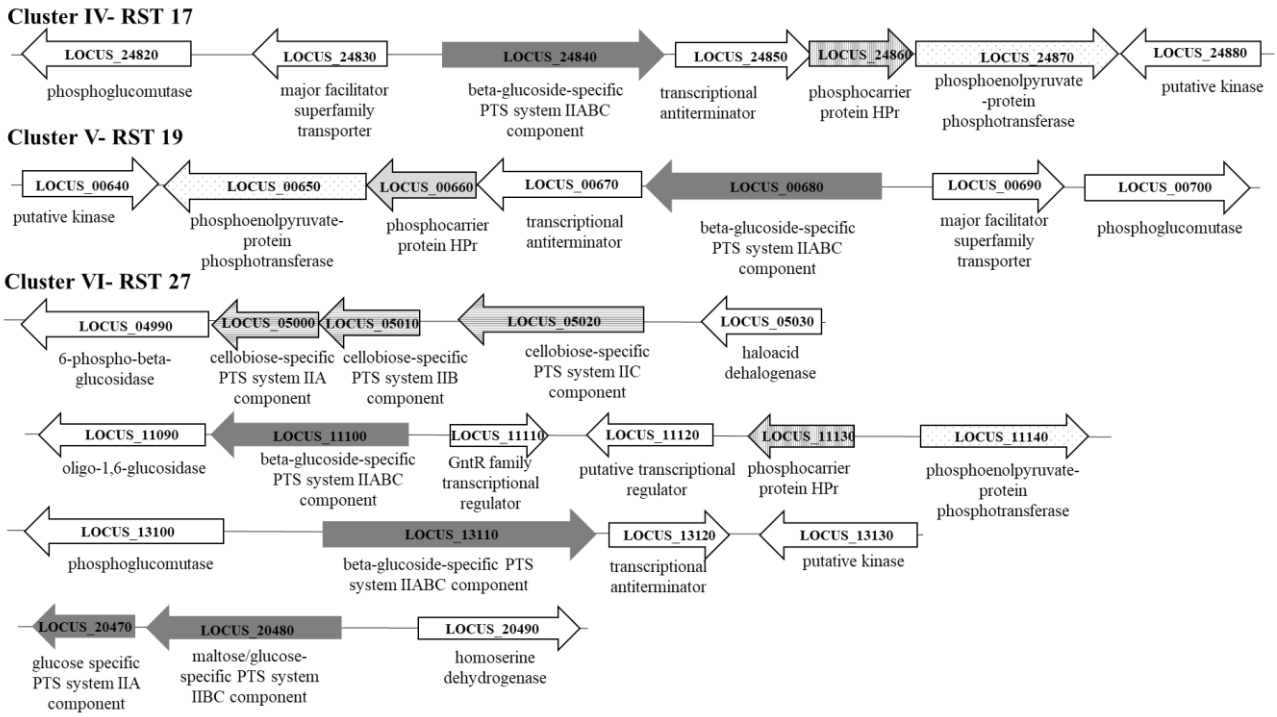


Cluster II- RST 8 & RST 16



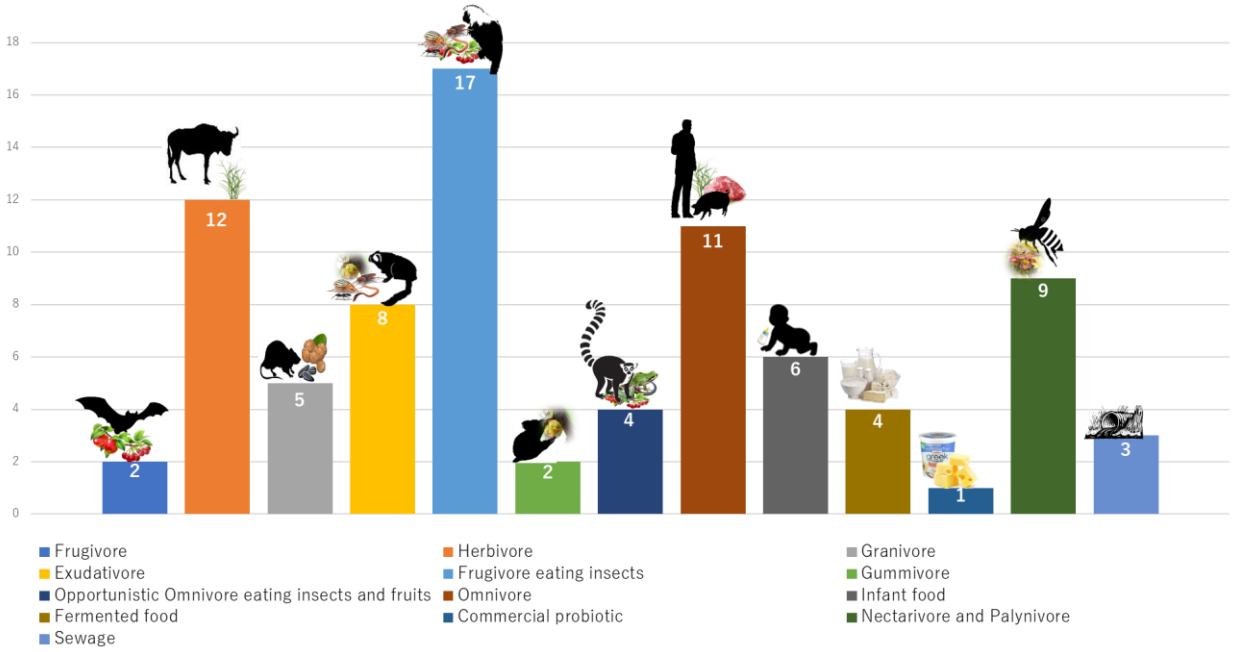
Cluster III- RST 9



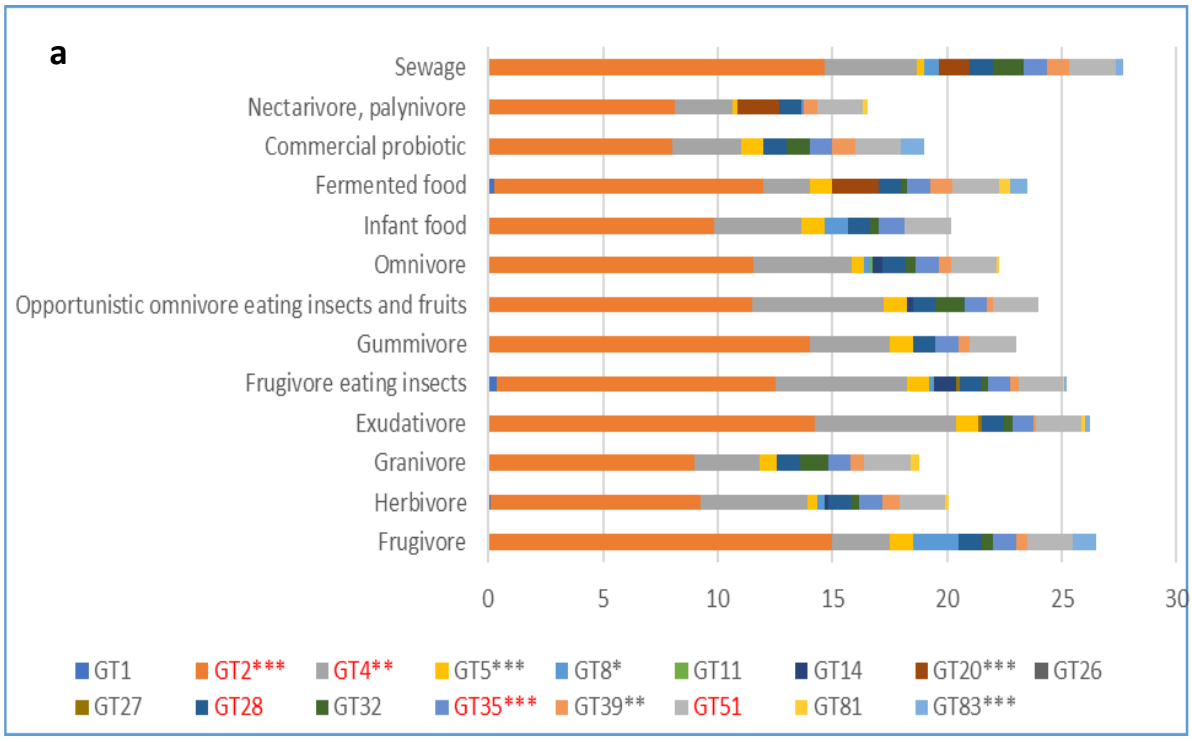


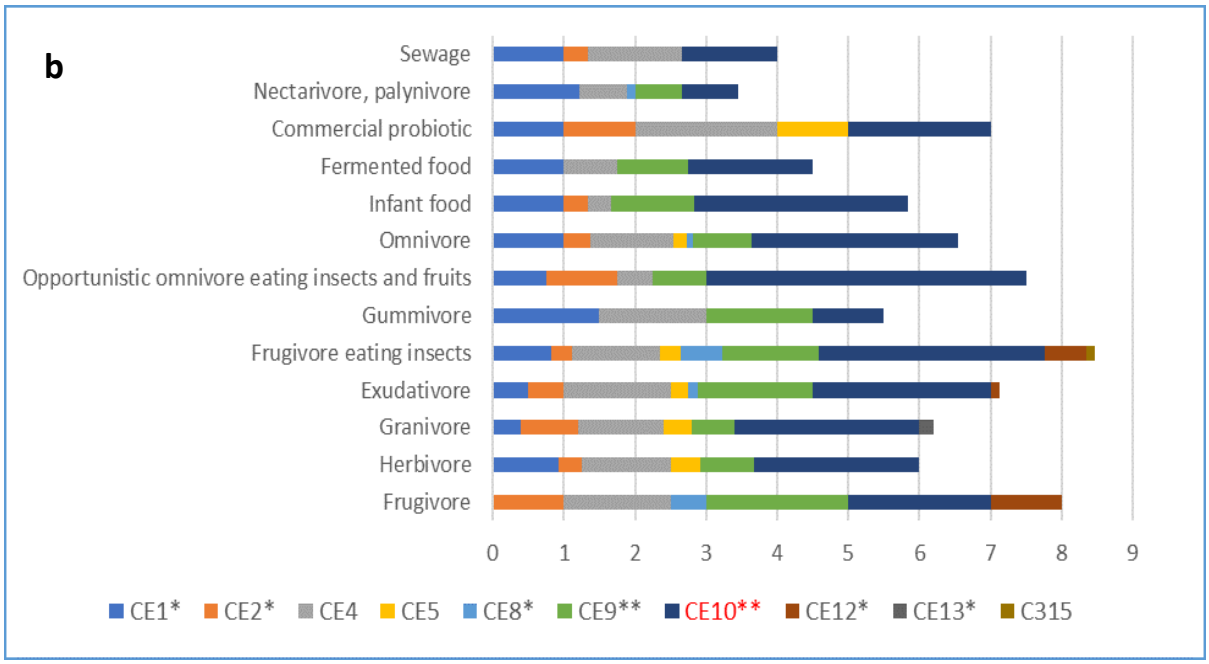
Supplementary Figure 4.2. Genetic maps of the predicted phosphotransferase system (PTS) gene clusters in genome of bat isolated bifidobacterial species.

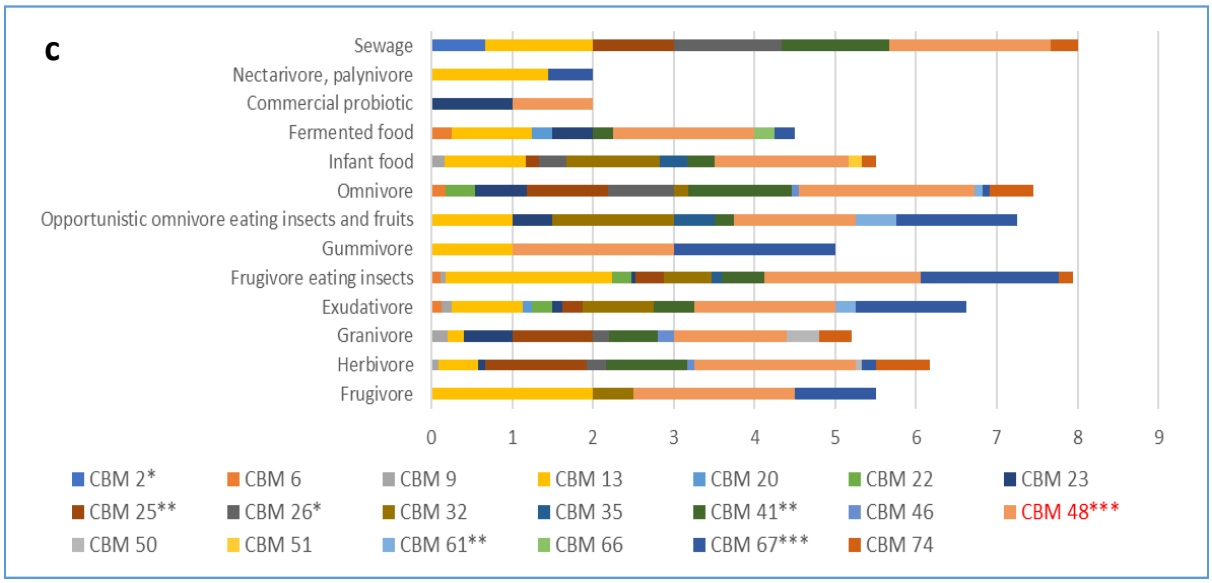
APPENDIX 3. SUPPLEMENTARY MATERIAL FOR CHAPTER 5



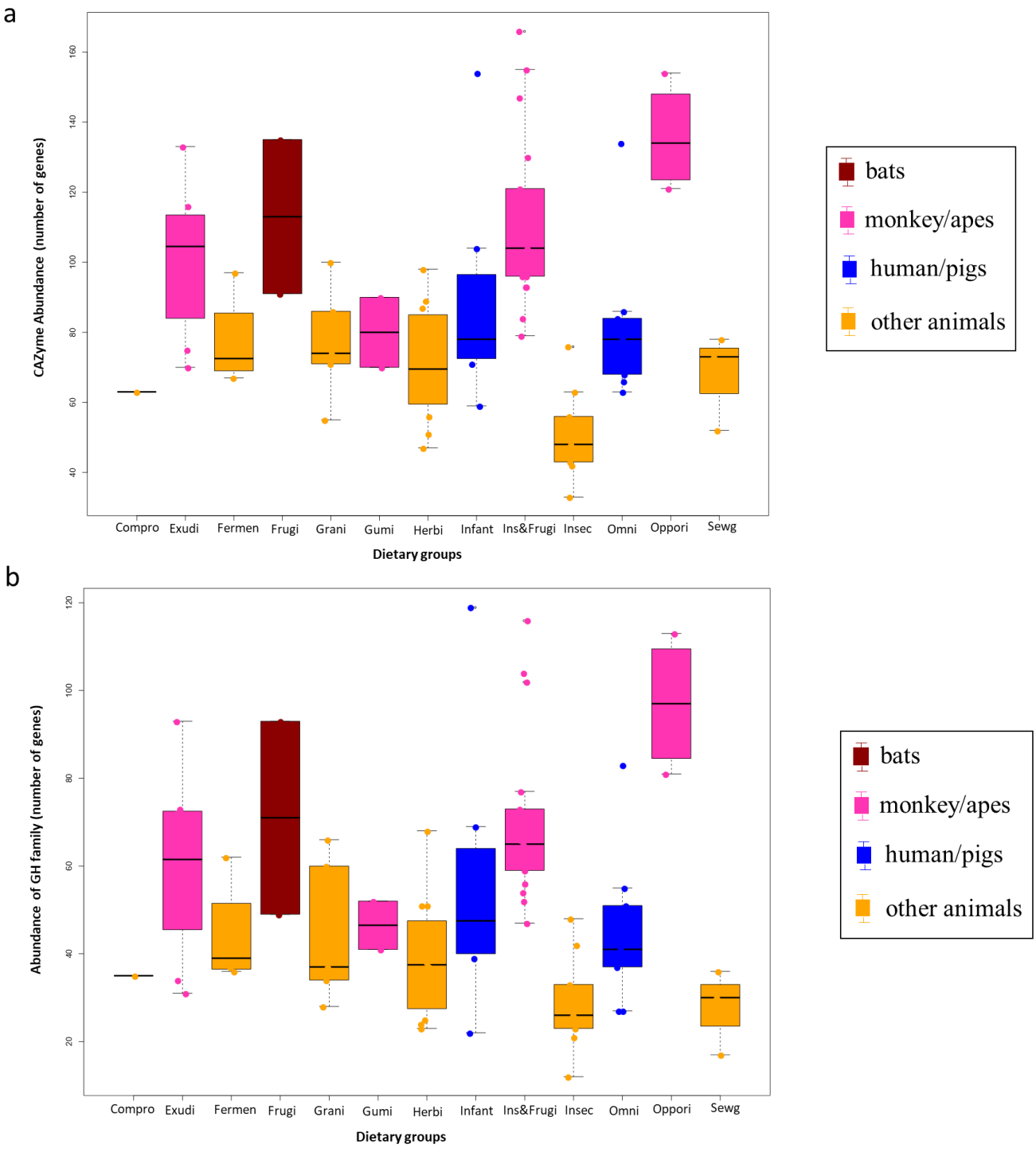
Supplementary Figure 5.1. Grouping of bifidobacterial species in accordance with the dietary pattern of their respective host, the chart displays the number of species belonging to each group [46].



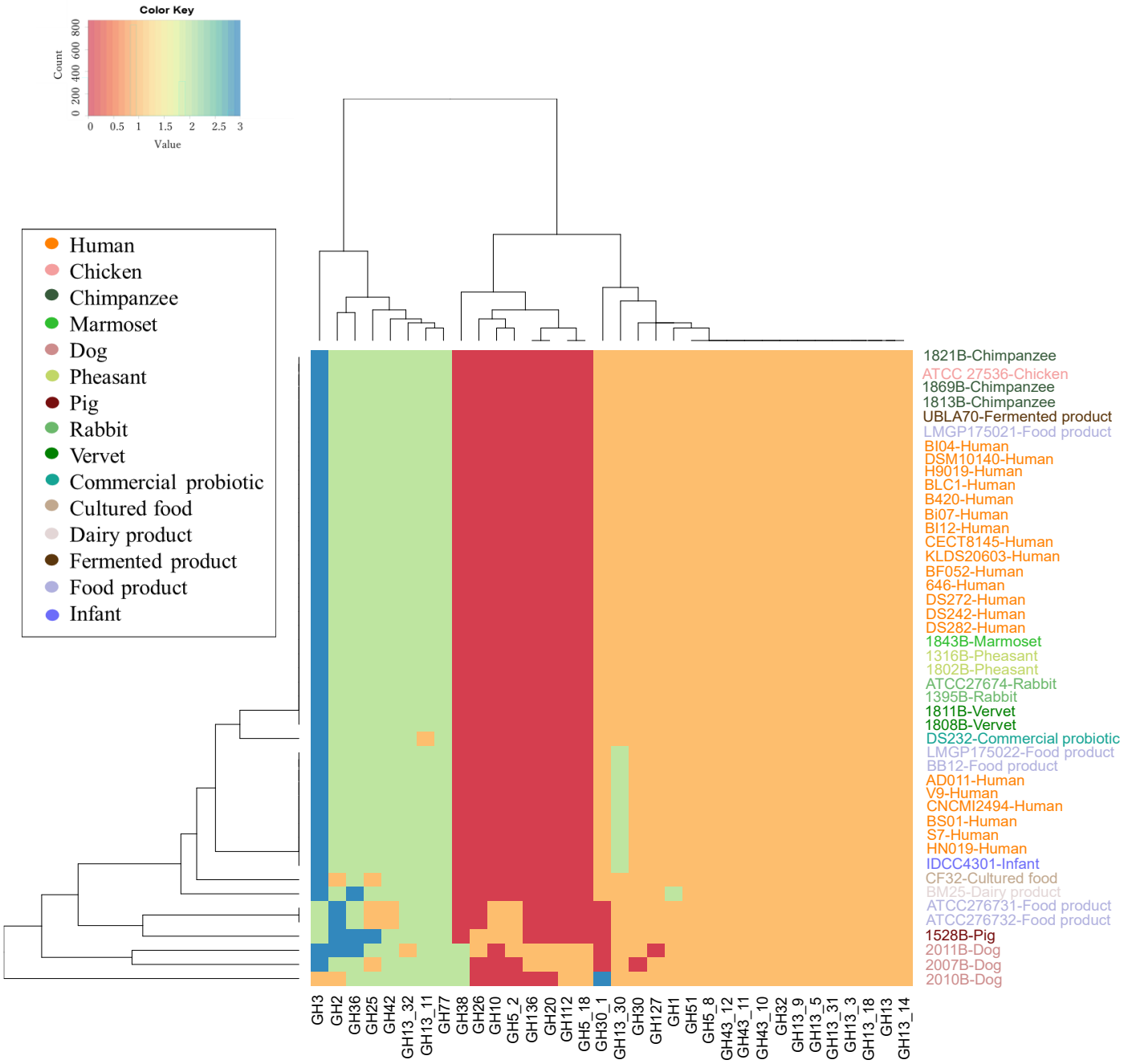




Supplementary Figure 5.2. Distribution of abundances of active carbohydrate enzyme family's genes in the dietary groups. (**a**) Glycosyltransferases (GT) family genes; (**b**) Carbohydrate esterase (CE) family genes; (**c**) Carbohydrate-binding module (CBM) family genes. Major CAZyme families present in more than 80% of the strains are highlighted in red color. The significance is shown by asterisks. *p <0.05, **p<0.01, ***p<0.001 [46].



Supplementary Figure 5.3. CAZyme families and GHs genes encoded by the strains in each dietary group **(a)** Abundances of the CAZyme families. **(b)** Abundance of genes encoding GHs. Compro: Commercial probiotic; Exudi: Exudativore; Fermen: Fermented food; Frugi: Frugivore; Grani: Granivore; Gumi: Gummivore; Herbi: Herbivore; Infant: Infant food; Ins&Frugi: Frugivore eating insects; Insec: Nectarivore and palynivore; Omni: Omnivore; Oppori: Opportunistic omnivore eating fruits, leaves and insects; Sewg: Sewage [46].



Supplementary Figure 5.4. Clustering of *B. animalis subsp. lactis* strains isolated from different isolation sources. Heatmap of strains based on gene count for GHs. Strains are colored according to their isolation source as represented in the legend [46].

Supplementary Table 5.1. Genomic features and diet information of all type strains [46].

Isolation Sources	Dietary Groups	Species	Specific host	Strains	GenBank Accession	Assembly Level	Genome Size Mb	GC Content %	CDS
Bat	Frugivore	*B. vespertilionis*	Egyptian fruit-bat	DSM 106025	RZOA00000000.1	Scaffold	3.075992	64.2	2409
		B. rousetti	Egyptian fruit-bat	DSM 106027	PEBH00000000.1	Scaffold	3.053799	64.6	2593
Geese	Herbivore	*B. anseris*	Domestic goose	LMG 30189	NMYC00000000.1	Contig	2.166761	64.3	1718
Rabbit		*B. cuniculi*	Rabbit feces	LMG 10738	JGYV00000000.1	Contig	2.531592	64.9	2194
		B. magnum	Rabbit feces	LMG 11591	JGZB00000000.1	Contig	1.822476	58.7	1507
		B. pullorum subsp. *saeculare*	Rabbit feces	LMG 14934	JGZM00000000.1	Contig	2.263283	63.7	1857
		B. italicum	European rabbit	LMG 30187	MVOG00000000.1	Contig	2.276351	65.4	1772
Bovine rumen		*B. merycicum*	Rumens of Cattle	LMG 11341	JGZC00000000.1	Contig	2.280234	60.3	1740
		B. pseudolongum subsp. *globosum*	Bovine rumen	DSM 20092	CP017695.1	Contig	1.935255	63.4	1574
		B. ruminantium	Rumens of Cattle	LMG 21811	JGZL00000000.1	Contig	2.249807	59.2	1832
		B. boum	Rumens of Cattle	DSM 20432	JHWO00000000.1	Contig	2.164426	52.8	1726
Sloth		*B. cholopei*	Sloth	BRDM6	VYSG00000000.1	Scaffold	2.248659	65	1667
Rodent		*B. castoris*	European beaver	LMG 30937	QXGI00000000.1	Contig	2.496067	65.4	2064
		B. dolichotidis	Patagonian mara	LMG 30941	QXGM00000000.1	Contig	1.921709	50.4	1452
	Granivore	*B. animalis* subsp. *animalis*	Mouse	ATCC 25527	CP002567.1	Complete	1.932693	60.5	1538
		B. tsurumiense	Golden hamsters	JCM 13495	JGZU00000000.1	Contig	2.164426	52.8	1629
		B. criceti	European hamster	LMG 30188	MVOH00000000.1	Contig	2.155882	62.5	1727
Chicken		*B. pullorum* subsp. *gallinarum*	Chicken cecum	LMG 11586	JGYX00000000.1	Contig	2.160836	64.2	1654
		B. pullorum subsp. *pullorum*	Chicken	DSM 20433	JDUI00000000.1	Contig	2.153559	64.2	1691
Non Human Primates	Exudativore	*B. tissieri*	Common marmosets	DSM 100201	MWWV00000000.1	Contig	2.873483	61.1	2235
		B .reuteri	Common marmosets	DSM 23975	JGZK00000000.1	Contig	2.847572	60.5	2149
		B. jachii	Common marmosets	DSM 103362	RQSP00000000.1	Scaffold	2.877198	62.2	2040
		B. myosotis	Common marmosets	DSM 100196	MWWW00000000.1	Contig	2.944195	62.6	2135
		B. catulorum	Common marmosets	DSM 103154	QFFN00000000.1	Scaffold	2.611484	63.2	2041
		B. callitrichos	Common marmosets	DSM 23973	JGYS00000000.1	Contig	2.887313	63.5	2364
		B. hapali	Common marmosets	DSM 100202	MWWY00000000.1	Contig	2.834308	54.5	2136
		B. aesculapii	Common marmosets	DSM 26737	BCFK00000000.1	Contig	2.693486	64.8	1924
	Frugivore eating insects	*B. vansinderenii*	Emperor tamarin	LMG 30126	NEWD00000000.1	Contig	3.111005	62.5	2497
		B. imperatoris	Emperor tamarin	LMG 30297	NMWV00000000.1	Contig	2.639899	56.1	2160
		B. callitrichidarum	Emperor tamarin	DSM 103152	QFFM00000000.1	Scaffold	3.121265	61.8	2548
		B. simiarum	Cotton top tamarin and emperor tamarin	DSM 103153	PEBK00000000.1	Contig	2.721281	63.8	2118
		B. scaligerum	Cotton top tamarin and	DSM 103140	PGLQ00000000.1	Scaffold	2.652159	58.3	2091

			emperor tamarin						
		B. primatium	Cotton top tamarin and emperor tamarin	DSM 100687	PEBI00000000.1	Scaffold	2.696768	63.2	2105
		B. felsineum	Cotton top tamarin and emperor tamarin	DSM 103139	PEBJ00000000.1	Scaffold	2.384744	57.1	1893
		B. callimiconis	Goeldi's marmoset	LMG 30938	QXGJ00000000.1	Contig	2.962404	62.4	2297
		B. goeldii	Goeldi's marmoset	LMG 30939	QXGL00000000.1	Contig	2.607372	56.1	2055
		B. stellenboschense	Red-handed tamarind	DSM 23968	JGZP00000000.1	Contig	2.812864	65.3	2203
		B. saguini	Red-handed tamarind	DSM 23967	JGZN00000000.1	Contig	2.787036	56.4	2321
		B. biavatii	Red-handed tamarind	DSM 23969	JGYN00000000.1	Contig	3.252147	63.1	2557
		B. platyrrhinorum	Squirrel monkey	DSM 106029	WHZV00000000.1	Scaffold	2.282466	62.6	1953
		B. ramosum	Cotton top tamarin	DSM 100688	WBSM00000000.1	Contig	3.046006	63.5	2334
		B. avesanii	Cotton top tamarin	DSM 100685	WBSN00000000.1	Contig	2.682617	66.3	2091
		B. aerophilum	Cotton top tamarin	DSM 100689	WHZW00000000.1	Scaffold	3.000921	63.6	2335
		B. simiasciurei	Squirrel monkey	DSM 106020	WHZU00000000.1	Scaffold	2.762496	63.6	2141
	Gummivore	*B. parmae*	Pygmy marmoset	LMG 30295	NMWT00000000.1	Contig	2.820211	65.8	2237
		B. margollesii	Pygmy marmoset	LMG 30296	NMWU00000000.1	Contig	2.789387	61.9	2221
	Opportunistic omnivore eating insects and fruits	*B. lemurum*	Ring-tailed lemur	DSM 28807	MWWX00000000.1	Contig	2.912024	62.6	2211
		B. samirii	Black-capped squirrel monkey	LMG 30940	QXGK00000000.1	Contig	2.574625	66.6	2013
		B. eulemuris	Adult black lemurs	DSM 100216	MWWZ00000000.1	Contig	2.913389	62.2	2315
		B. moukalabense	Wild lowland gorilla	DSM 27321	AZMV00000000.1	Contig	2.515335	59.9	2046
Pig	Omnivore	*B. choerinum*	Piglet faeces	LMG 10510	JGYU00000000.1	Contig	2.096123	65.5	1672
		B. pseudolongum subsp. *pseudolongum*	Pig faeces	LMG 11571	JGZH00000000.1	Contig	1.898684	63.1	1495
		B. longum subsp. *suis*	Pig faeces	DSM 20211	JDUC00000000.1	Scaffold	2.602875	59.9	2032
		B. thermacidophilum subsp. *porcinum*	Piglet faeces	DSM 17755	JDTQ01000001.1	Contig	2.079368	60.2	1738
		B. thermophilum	Human Adult	JCM 1207	JGZV00000000.1	Complete	2.291643	60.1	1845
Human Adult		*B. adolescentis*	Human Adult	ATCC 15703	AP009256.1	Complete	2.089645	59.2	1631
		B. angulatum	Human Adult	DSM 20098	AP012322.1	Complete	2.021974	59.4	1585
		B. dentium	Human Adult	LMG 20436	AP012326.1	Complete	2.635669	58.5	2141
		B. gallicum	Human Adult	LMG 11596	JGYW00000000.1	Contig	2.004594	57.6	1507
		B. longum subsp. *longum*	Human Adult	JCM 1217	AP010888.1	Complete	2.385164	60.3	1924
		B. scardovii	Human Adult	DSM 13734	AP012331.1	Complete	3.158347	64.6	2572
Human Infant	Infant food	*B. bifidum*	Human Infant	ATCC 29521	AP012323.1	Complete	2.214656	62.7	1707
		B. breve	Human Infant	DSM 20213	AP012324.1	Complete	2.269415	58.9	1929

		B. catenulatum subsp. *catenulatum*	Human Infant	LMG 11043	AP012325.1	Complete	2.079525	56.2	1710
		B. catenulatum subsp. *kashiwanohense*	Human Infant	DSM 21854	AP012327.1	Complete	2.337234	56.3	1945
		B. longum subsp. *infantis*	Human Infant	DSM 20088	CP001095.1	Complete	2.832748	59.9	2416
		B. pseudocatenulatum	Human Infant	LMG 10505	AP012330.1	Complete	2.313752	56.4	1841
Fermented Products	Fermented food	*B. aquikefiry*	Fermented Products	LMG 28769	MWXA00000000.1	Contig	2.408364	52.3	1982
		B. crudilactis	Fermented Products	LMG 23609	JHAL00000000.1	Contig	2.362816	57.7	1800
		B. mongoliense	Fermented Products	DSM 21395	JGZE00000000.1	Contig	2.17049	62.8	1798
		B. psycraerophilum	Fermented Products	LMG 21775	JGZI00000000.1	Contig	2.615078	58.8	2122
Commercially used probiotic	Commercial probiotic	*B. animalis* subsp. *lactis*	Commercial probiotic	DSM 10140	CP001606.1	Complete	1.938483	60.5	1566
Bees	Nectarivore and palynivore	*B. actinocoloniiforme*	Bumble bees	DSM 22766	CP011786.1	Complete	1.83006	62.7	1296
		B. asteroides	Honey bees	DSM 20089	CP017696.1	Complete	2.167304	60.1	1659
		B. bohemicum	Bumble bees	DSM 22767	JGYP00000000.1	Contig	2.05247	57.5	1632
		B. bombi	Bumble bees	DSM 19703	ATLK00000000.1	Contig	1.895239	56.1	1454
		B. coryneforme	Honey bees	LMG18911	CP007287.1	Complete	1.755151	60.5	1364
		B. commune	Bumble bees	DSM 28792	FMBL00000000.1	Scaffold	1.633662	53.9	1238
		B. indicum	Honey bees	LMG 11587	CP006018.1	Complete	1.734546	60.5	1350
		B. xylocopae	Carpenter bee	LMG 30142	PDCH00000000.1	Contig	1.848461	62.8	1476
		B. aemilianum	Carpenter bee	LMG 30143	PDCG00000000.1	Contig	2.017578	61.1	1640
Sewage	Sewage	*B. minimum*	Sewage	LMG 11592	JGZD00000000.1	Contig	1.89286	62.7	1590
		B. subtile	Sewage	LMG 11597	JGZR00000000.1	Contig	2.790088	60.9	2260
		B. thermacidophilum subsp. *thermacidophilum*	Sewage	LMG 15837	AUFI00000000.1	Contig	2.233072	60.4	1824

Supplementary Table 5.2. Isolation sources and accession numbers for 66 strains isolated from various hosts [46].

Serial Number	Specie Name	Strain Name	Isolation Source	GenBank Accession
1	*B. animalis* subsp. *lactis*	CNCM I2494	Dairy product	CP002915.1
2	*B. animalis* subsp. *lactis*	BLC1	Probiotic	CP003039.1
3	*B. animalis* subsp. *lactis*	B420	Human	CP003497.1
4	*B. animalis* subsp. *lactis*	Bl12	Human	CP004053.1
5	*B. animalis* subsp. *lactis*	ATCC27673	Sewage	CP003941.1
6	*B. animalis* subsp. *lactis*	ATCC27536	Chicken	AWFL00000000.1
7	*B. animalis* subsp. *animalis*	IM386	Human	CBUQ000000000.1
8	*B. animalis* subsp. *animalis*	MCC1489	Pig	AWFO00000000.1

9	*B. animalis* subsp. *animalis*	MCC0483	Rodent	AWFK00000000.1
10	*B. animalis* subsp. *animalis*	ATCC27672	Rodent	AWFQ00000000.1
11	*B. animalis* subsp. *animalis*	YL2	Rodent	NHMR00000000.2
12	*B. longum* subsp. *infantis*	BIC1307292462	Probiotic	CCWO00000000.1
13	*B. longum* subsp. *infantis*	BIC1401212621b	Probiotic	CCWS00000000.1
14	*B. longum* subsp. *infantis*	BT1	Infant	CP010411.1
15	*B. longum* subsp. *infantis*	CECT 7210	Infant	CELR00000000.1
16	*B. longum* subsp. *suis*	VT155	Calf	SAMN17849156
17	*B. longum* subsp. *suis*	VT247	Calf	SAMN17849157
18	*B. longum* subsp. *suis*	SU851	Pig	WHVJ00000000
19	*B. longum* subsp. *suis*	LMG 21814	Pig	JGZA00000000.1
20	*B. longum* subsp. *suis*	DSM 20211	Pig	JDUC00000000.1
21	*B. longum* subsp. *suis*	BSM11-5	Infant	MOAE00000000.1
22	*B. breve*	B7212	Human	SAMN17849159
23	*B. breve*	B2150	Infant	SAMN17849160
24	*B. breve*	JCM 7019	Human	CP006713.1
25	*B. breve*	ACS-71-V-Sch8b	Human	CP002743.1
26	*B. breve*	CECT 7263	Human milk	AFVV00000000.1
27	*B. breve*	S27	Infant	CP006716.1
28	*B. breve*	DSM 20213	Infant	ACCG00000000.2
29	*B. breve*	UCC2003	Infant	CP000303.1
30	*B. bifidum*	B7298	Human	SAMN17849161
31	*B. bifidum*	B2662	Infant	SAMN17849162
32	*B. bifidum*	VT188	Calf	SAMN17849163
33	*B. bifidum*	B2009	Infant	SAMN17849164
34	*B. bifidum*	B7313	Human	SAMN17849165
35	*B. thermophilum*	RBL67	Human	CP004346.1
36	*B. thermophilum*	DSM20212	Bovine	JHWM00000000.1
37	*B. thermophilum*	JCM1207	Pig	JGZV00000000.1
38	*B. thermophilum*	DSM20210	Pig	JDUB00000000.1
39	*B. thermophilum*	1543B	Pig	PCGX00000000.1
40	*B. thermophilum*	1542B	Pig	PCGY00000000.1
41	*B. adolescentis*	22L	Human	CP007443.1
42	*B. adolescentis*	BBMN23	Human	CP010437.1
43	*B. adolescentis*	LMG11579	Bos taurus	LNKL00000000.1
44	*B. adolescentis*	UBA2084	Sewage	DCZM00000000.1
45	*B. pseudocatenulatum*	IPLA36007	Human	JEOD00000000.1
46	*B. pseudocatenulatum*	1E	Calf	MNLB00000000.1
47	*B. pseudocatenulatum*	12	Human	CP025199.1
48	*B. pseudolongum* subsp. *globosum*	LMG11569	Bos taurus	JGZG00000000.1
49	*B. pseudolongum* subsp. *globosum*	DSM20092	Bovine	CP017695.1
50	*B. pseudolongum* subsp. *globosum*	1744B	Bear	PCHB00000000.1
51	*B. pseudolongum* subsp. *globosum*	1619B	Llama	PCHC00000000.1
52	*B. pseudolongum* subsp. *globosum*	1549B	Chicken	PCHG00000000.1

53	*B. pseudolongum* subsp. *globosum*	1520B	Rodent	PCHH00000000.1
54	*B. pseudolongum* subsp. *globosum*	1524B	Rodent	PCGZ00000000.1
55	*B. pseudolongum* subsp. *globosum*	1747B	Giraffe	PCHA00000000.1
56	*B. pseudolongum* subsp. *globosum*	1734B	Wallaby	PCHD00000000.1
57	*B. pseudolongum* subsp. *globosum*	1691B	Hippopotamus	PCHE00000000.1
58	*B. pseudolongum* subsp. *pseudolongum*	1370B	Pig	PCHI00000000.1
59	*B. pseudolongum* subsp. *pseudolongum*	2054B	Dog	RYUN00000000.1
60	*B. pseudolongum* subsp. *pseudolongum*	1629B	Tapir	RYVE00000000.1
61	*B. moukalabense*	GG01	Western Lowland Gorilla	BJEZ00000000.1
62	*B. moukalabense*	GB01	Western Lowland Gorilla	BJFA00000000.1
63	*B. moukalabense*	CD14	Chimpanzee central	BJFG00000000.1
64	*B. moukalabense*	CD16	Chimpanzee central	BJFH00000000.1
65	*B. moukalabense*	EB43	African forest elephant	BJFJ00000000.1
66	*B. moukalabense*	EB44	African forest elephant	BJFK00000000.1

Supplementary Table 5.3. Isolation sources and accession numbers for 45 *B. animalis* subsp. *lactis* strains [46].

Strain Name	**Isolation Source**	**GenBank Accession**
ATCC 27536	Chicken	AWFL00000000.1
1821B	Chimpanzee	RSCT00000000.1
1869B	Chimpanzee	RSCR00000000.1
1813B	Chimpanzee	RSCU00000000.1
DS23_2	Commercial dietary supplements	QDIO00000000.1
CF3_2	Cultured Food	QDIV00000000.1
BM 25	dairy product	PHUS00000000.1
2010B	Dog	RSCP00000000.1
2011B	Dog	RSCO00000000.1
2007B	Dog	RSCQ00000000.1
UBBLa 70	fermented food	RWKO00000000.1
BB-12	Food product	PESQ00000000.1
ATCC 27673	Food product	CP003941.1
ATCC 27673	Food product	AWFP00000000.1
LMG P-17502_1	Food product	NIGR00000000.1
LMG P-17502_2	Food product	NIGQ00000000.1
AD011	Human	CP001213.1
Bl-04; ATCC SD5219	Human	CP001515.1

DSM 10140	Human	CP001606.1
V9	Human	CP001892.1
HN019	Human	ABOT00000000.1
CNCM I-2494	Human	CP002915.1
BLC1	Human	CP003039.1
BS 01	Human	AHGW00000000.1
B420	Human	CP003497.1
Bi-07	Human	CP003498.1
Bl12	Human	CP004053.1
CECT 8145	Human	CBWX000000000.1
KLDS2.0603	Human	CP007522.1
BF052	Human	CP009045.1
S646	Human	MLZL00000000.1
DS27_2	Human	QDIL00000000.1
DS24_2	Human	QDIN00000000.1
DS28_2	Human	QDIK00000000.1
S7	Human	CP022724.1
HN019	Human	CP031154.1
IDCC4301	Infant	CP031703.1
1843B	Marmoset	RSCS00000000.1
1316B	Pheasant	RSDA00000000.1
1802B	Pheasant	RSCX00000000.1
1528B	Pig	RSCY00000000.1
ATCC 27674	Rabbit	AWFM00000000.1
1395B	Rabbit	RSCZ00000000.1
1811B	Vervet	RSCV00000000.1
1808B	Vervet	RSCW00000000.1