UNIVERSITÀ
DI PARMA

UNIVERSITÀ DEGLI STUDI DI PARMA

Dottorato di ricerca in biotecnologie e bioscienze

XXXVI ciclo

Engraftment, development, and roles of the infant gut microbiota

Coordinatore:
Chiar.mo Prof. Marco Ventura

Tutor:
Chiar.mo Prof. Christian Milani

Dottorando:

Chiara Tarracchini

Parma, 2020/2021-2022/2023

Table of Contents

Summary

Based on the widely accepted sterile womb paradigm, the fetal environment can be regarded as nearly sterile, and the first microbial colonization of the newborn's gut is believed to occur during delivery and shortly after birth through a combination of vertical transmission from the mother and horizontal acquisition from other humans and the environment. Following the first seeding, the gut microbiota evolves substantially, reaching complete maturation within the first three years of life. However, this process is influenced and can be disturbed by several external factors, such as maternal diet, gestational age, delivery mode, feeding type, and antibiotic use.

Given that inadequate gut microbiota development in early life is frequently associated with neonatal and long-term adverse health conditions, understanding the processes that govern initial colonization and development of the infant gut microbial community is of great importance.

The aim of this Ph.D. thesis is to explore the engraftment and evolution of the infant gut microbiota by exploiting the most reliable genomic, metagenomic, and phylogenomic approaches. Within this framework, strain-level tracking of gut commensals revealed that beneficial, maternally inherited bifidobacterial species constitute highly stable and resilient communities, identifying host's sex as a potential variable affecting the long-term persistence of these microorganisms through infancy. In addition to developmental changes in composition, this Ph.D. thesis also explores the functional maturation of the healthy human gut microbiota by tracing the microbial genetic potential for bioactive metabolites from infancy to adulthood, specifically emphasizing the early stages of life.

As vaginal-derived microbes can be implicated in the first seeding of the infant gut microbiota, one of the purposes of this Ph.D. thesis is to unravel the composition of

the vaginal microbiota across the population, performing detailed genome comparative analyses of *Lactobacillus crispatus* and *Gardnerella vaginalis* species, which are notoriously associated with vaginal health and potential dysbiotic status, respectively.

As mentioned above, various perinatal factors can significantly impact the developmental trajectories of the nascent gut-associated microbial community. Among these, gestational age at birth is regarded as one of the most impactful, as it dictates the degree of immaturity of several infant's organs, including intestinal and immune systems. Consistently, premature birth and low birth weight (< 1,500 gr) have been associated with an increased risk of Necrotizing EnteroColitis (NEC). In this context, this Ph.D. thesis explores the gut microbiota composition of preterm infants affected by NEC and prior to NEC onset to identify possible early microbial and functional biomarkers of this severe disease.

Given the broadly observed link between early-life depletion of *Bifidobacterium* genus and host diseases, members of this genus are increasingly used as potential players in restoring gastrointestinal functions. Accordingly, in this Ph.D. thesis, the genetic traits involved in growth, microbe-host, and microbe-microbe interactions of bifidobacterial strains were explored, with a particular focus on *B. longum* subsp. *infantis* members. Moreover, common bifidobacterial genomes used in commercial probiotic products were used to build a free-access genomic database named Integrated Probiotic DataBase (IPDB).

Finally, considering the bidirectional association between host's factors and the gut microbiota composition, in this Ph.D. thesis, we determine the impact of different human lifestyles, i.e., sedentary or athletic, on the gut microbiota composition, showing that the host and the bacterial community inhabiting the gut are continuously crosstalking, mutually modulating each other throughout lifespan.

Chapter 1

General Introduction

A. The origin of the first microbial colonizer of the neonatal gut

The "sterile womb paradigm"

In the last century, it has been assumed that the intrauterine environment in healthy pregnancies is sterile. However, this dogma has been challenged in the past decade, particularly with the introduction of new molecular approaches. Indeed, studies utilizing a combination of next-generation sequencing techniques, quantitative PCR (qPCR), and fluorescence in situ hybridization (FISH) suggested the existence of a specific microbiome in the placenta, amniotic fluid, fetal lung, and meconium[1–5], offering an original perspective on early microbiome development. Nevertheless, the concept of a fetal microbiome has not yet gained widespread acceptance within the scientific community[6–10]. A central matter of discussion revolves around discerning the presence of microbial genetic material from the existence of active, metabolically functioning microorganisms[7,11,12].

Interestingly, recent data have suggested that bacterial DNA within the neonatal meconium may be transferred from the mother to the developing fetus during pregnancy, potentially conveyed in extracellular vesicles renowned for their ability to cross biological barriers, including the placenta[1]. Moreover, it has been argued that the composition of the low bacterial biomass identified in placenta and meconium samples is easily influenced by environmental exposures and post-birth factors[9,13]. Notably, the composition of the microbiota detected in the first stool after birth, formed in utero, appeared to be affected by the mode of delivery, corroborating the potential perinatal colonization[1,14]. In general, the debate over the existence of an in-utero microbiota is fueled by several issues, and the existence of a fetal microbiota remains under discussion.

Maternal inheritance of gut bacteria

In line with the widely recognized sterile womb paradigm, the initial formation of the infant gut microbiota is a multifaceted process influenced by several factors, with a key determinant being childbirth. Specifically, the first microbial colonization of the neonatal intestine is believed to occur during delivery (or following rupture of the amniotic membranes) and shortly after birth through a combination of vertical transmission from the mother and horizontal acquisition from other humans or the environment.

Several lines of evidence have revealed that a substantial fraction (ranging from 30 to 70%) of the microbial species in the infant's gut on the day of delivery was transmitted from the gut, vaginal tract, oral cavity, or skin of the mother, remaining relatively stable over the following months[15,16] (Figure 1). Indeed, a recent investigation based on single nucleotide variant profiles showed that most bacterial strains found in newborns acquired from their mothers tend to persist in the infant's gut for an extended period, while microorganisms from nonmaternal sources tend to be replaced within the first year of age[16,17]. This observation points to a natural selection process favoring specific maternal-derived microbes demonstrating high ecological adaptability to colonize the infant's gastrointestinal tract.

Consistently with recent strain-level metagenomic analyses, the gut microbiota is the prominent maternal source of infant-inherited bacteria, followed by the vaginal, oral, and skin microbiomes[16,18] (Figure 1). Specifically, early-life specialists such as *Bifidobacterium bifidum*, *Bifidobacterium breve*, and *Bifidobacterium longum* are among the most frequently maternally transmitted species. Important functional roles in newborn development have been attributed to these microorganisms, including educating the immature immune system, balancing inflammation processes, and secreting factors that improve host health and neurodevelopment[19–22]. The specific

adaptation of bifidobacterial species for the infant gut environment lies in their ability to metabolize the complex oligosaccharides naturally found in human breast milk that are indigestible to humans (see below)[23–25]. Along with *Bifidobacterium* species, some *Bacteroides* species, such as *Bacteroides fragilis*, *Bacteroides dorei*, and *Bacteroides vulgatus*, are shared with the mother's gut microbiota within the first days after birth[26–28].

In the days following childbirth, the proportion of microbes shared between the mother and the newborns steadily expanded (Figure 1), facilitated by the close physical contact and the continuing exchange of microbes that occurs through breastfeeding.

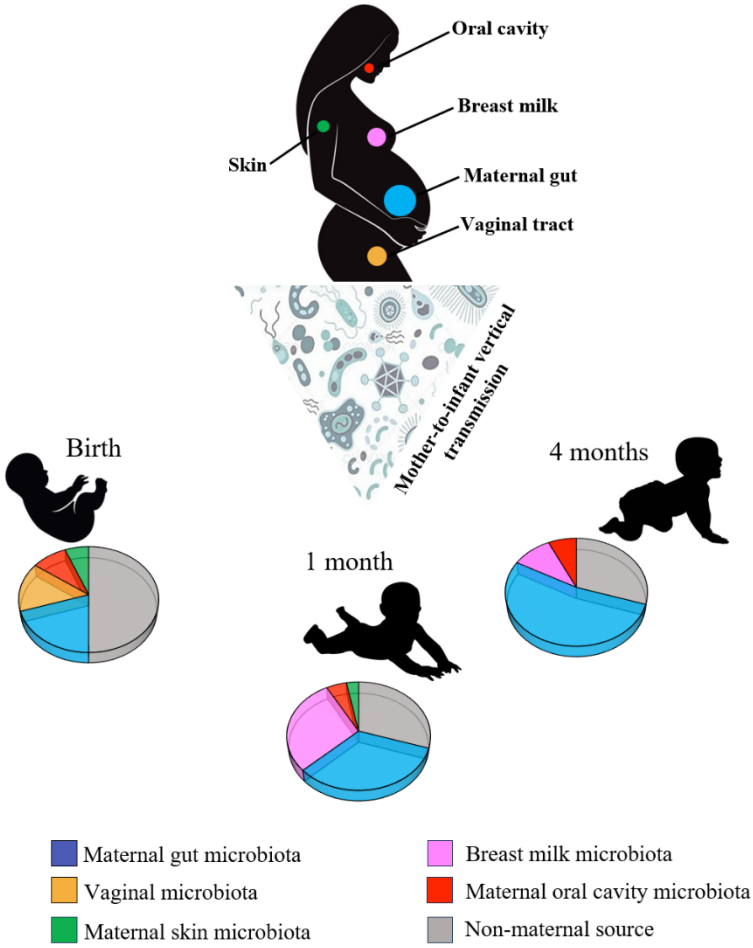


Figure 1. Schematic representation of the contribution of different maternal body sites on the first seeding of the infant's gut. Data concerning the contribution of maternal body sites are based on Ferretti et. al., 2018[16] and Pannaraj et. al., 2017[29].

However, maternal-derived microorganisms did not appear equally prone to stably colonize the infant's gut. For example, *Veillonella*, *Prevotella*, and *Streptococcus mitis* dominated the oral microbiota of pregnant women, but only *V. parvula* was found in the paired infant fecal samples[30,31]. Interestingly, it has been observed that, within the oral microbiota, *Veillonella* consumes lactate produced by *Streptococcus* or *Lactobacillus*, producing propionate and acetate as metabolic byproducts [32]. This trophic behavior indicates that the frequent presence of *Veillonella* in the infant's gut

microbiota may result from cooperation with lactic acid-producing bacteria that feed on the complex oligosaccharides in human milk. Contrarily, mother skin- and vaginal-derived bacteria like *Staphylococcus*, *Streptococcus*, *Propionibacterium*, and *Lactobacillus* are usually transient, not establishing long-term persistence in the infant's gut beyond the neonatal period (Figure 1), likely because of the differences in biotic and abiotic factors characterizing these ecological niches.

The importance of early maternal colonization is underscored by the observation that, in contrast to the skin, vagina, and oral cavity, most maternal gut-derived species tend to establish a long-lasting persistence in the infant gut[17]. This indicates unique compatibility between the maternal gut-derived bacteria and the infant, possibly supported by partially genetically determined factors, such as breastmilk oligosaccharide composition, intestinal mucus structure, and immune system functioning.

The vaginal microbiome as a source of microbes to the infant gut

To gain insights into the initial bacterial exposure that naturally occurs during vaginal childbirth, it is convenient to characterize the maternal vaginal microbiome. Among healthy non-pregnant women, researchers have identified five recurring microbial profiles, named community state types (CSTs)[33]. Four of these CSTs, i.e., CST I-III and V, are typically dominated by *Lactobacillus* species, including *Lactobacillus crispatus*, *Lactobacillus gasseri*, *Lactobacillus iners* and *Lactobacillus jensenii*, respectively, while the CST IV is characterized by modest abundance of *Lactobacillus* species, accompanied by several species of strictly anaerobic bacteria, such as *Streptococcus* and *Corynebacterium*, or members of *Gardnerella*, *Prevotella*, and *Atopobium* genera[34].

The presence of *Lactobacillus* species, which can account for more than 50% of the total vaginal commensals[33,35], is linked to a healthy state, and it is believed to protect the women's reproductive tract from the colonization by potential pathogens, mainly by producing hydrogen peroxide, bacteriocins, and lactic acid, which maintains a low vaginal pH (<4.5)[36,37]. Indeed, the most common form of vaginal microbiota dysbiosis affecting reproductive-aged women, i.e., bacterial vaginosis, is associated with the alteration of the vaginal microbiome from *Lactobacillus* dominance to a high abundance of anaerobic and facultative bacteria, such as *Gardnerella vaginalis*, with concomitant increased pH > 4.5 and greater microbial diversity[38,39]. Moreover, bacterial vaginosis has been linked to various reproductive tract disorders[40], including infertility, preterm labor and delivery, and susceptibility to viral infections[41,42]. For this reason, in the past decade, many efforts have been directed toward understanding this microbial community, its compositional changes throughout pregnancy, and its relationship with perinatal infections and premature birth.

Results from several recent studies showed that the cervicovaginal microbiome experiences several changes over the course of pregnancy, emphasizing the role of sex hormones in driving these transformations[43,44]. In this regard, culture-independent profiling methods revealed that the vaginal microbiota decreases biodiversity and increases stability until shortly before delivery, potentially lowering the risk of bacterial perturbations implicated in adverse pregnancy outcomes, including preterm delivery and low birth weight. The vaginal microbiome during a normal pregnancy is generally dominated by *L. crispatus*, which is responsible for lowering diversity and reducing the risk of colonization by diverse anaerobes and potential pathogens[45].

As mentioned above, the maternal vaginal microbiome can contribute to establishing the infant microbiota in early life, providing a natural first exposure of the newborn to microbes[46].

Consistently, it has been reported that a few days following birth, members of the vaginal microbiome constitute up to 20 % of the total gut microbiota of their own infants[47]. Specifically, the gut microbiota of vaginally-delivered infants at birth resembles the mother's vaginal microbiota, dominated by the *Lactobacillus* genus[47]. Nevertheless, as mentioned above, it has been found that most bacteria of vaginal origin do not persist for an extended period in the newborn's gut due to the distinct chemical, physical and biological factors differentiating the microaerophilic vaginal environment from that anaerobic of the gut. However, a recent study based on a mouse model showed that birth-associated exposure to maternal vaginal microbiota could have long-term effects on the offspring's health, modulating immune system development and activities[48].

Perinatal factors influencing the early-life gut microbiota

The development of the infant gut microbiome involves a *de novo* assembly of a highly intricate microbial community that establishes in a virtually empty niche. Accordingly, this process is highly vulnerable and can be influenced by maternal and infant-related factors, as well as environmental variables, acting in the perinatal (between five months before birth and one month after birth) and postnatal periods[49] (Figure 2).

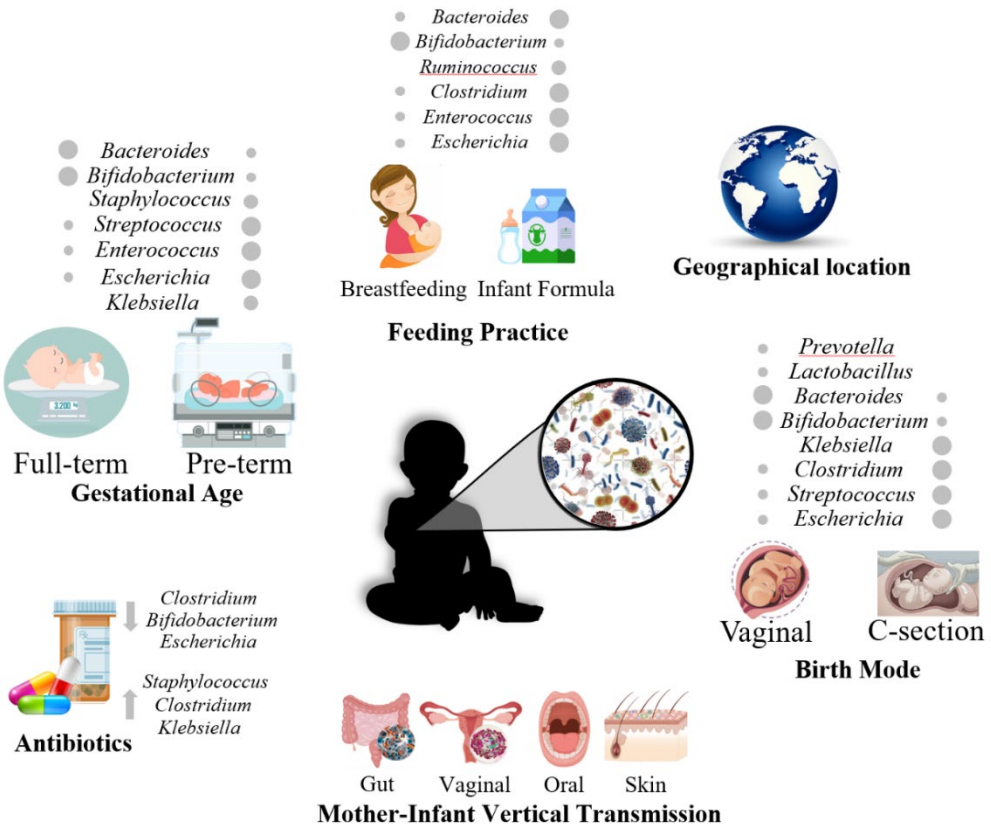


Figure 2. The main perinatal factors influencing the establishment and development of the infant gut microbiota. Differential compositional data of key microbial members are reproduced by Mancabelli et al. 2020[50].

As discussed above, the neonatal period represents a crucial early window of time for acquiring key gut commensals from maternal reservoirs, and disruption of this route can have profound consequences on the infant's gut microbiota structure. For example, considering different delivery modes (Cesarean section vs. vaginal delivery), it has been observed that in vaginally delivered newborns, more than 80% of the gut microbiota was shown to be derived from the mother, while infants born through Cesarean section (C-section) acquired only 12.56% of their early gut microbial community from maternal microbial reservoirs[51].

Specifically, infants born through C-section (without amniotic membrane rupture) exhibited a low presence of vaginal microbes (e.g., *Lactobacillus* and *Prevotella* spp.) in their gut microbiome at birth due to the absence of contact with the vaginal canal environment (Figure 2). Instead, they are more susceptible to being colonized by bacteria from the mother's skin and other humans (e.g., *Staphylococcus*, *Corynebacterium*, *Propionibacterium* spp.)[46,52], as well as by potential microbial pathogens encountered in hospital environments, such as *Clostridium*, *Streptococcus*, and *Klebsiella*[51,53] (Figure 2).

In addition, postnatal colonization of the neonatal intestine by *Bacteroides* and *Bifidobacterium* spp., responsible for HMO metabolism and often vertically transmitted from the mother, appeared delayed or completely absent in infants delivered by C-section[26,54] (Figure 2). Although differences in microbial composition introduced by delivery mode tend to normalize within the first years after birth[51], current research evidenced that shifts in the early microbial exposures can lead to long-term health consequences, including increased risk of chronic immune diseases, obesity, and allergic disorders[55–57].

While the delivery mode is perhaps regarded as the strongest perinatal factor modulating the first microbial colonization of the infant's gut, the feeding type can profoundly modulate the immediately following stages of life. Similar to the observations in vaginally delivered infants, several studies concur in emphasizing that, within six months, *Bifidobacterium* genus is more represented in exclusively breastfed infants compared to those receiving formula[50], who, in contrast, exhibit a more diverse microbial profile with a relatively higher abundance of *Bacteroides*, *Enterococcus*, *Ruminococcus*, *Staphylococcus*, *Escherichia*, and *Clostridium* genera[58,59] (Figure 2). This observation highlights that vertical transmission events of bacteria from mother to child, coupled with intrinsic selective factors in breast

milk, synergistically ensure the persistence of key microbial taxa in the infant's intestinal environment. This natural process promotes the proper formation of the gut microbiota and prevents various infectious morbidities, eventually playing a crucial role in the overall development and long-term well-being of the infant.

Along with delivery mode and feeding type, gestational age at birth is another important perinatal factor influencing the establishment of the infant gut microbiota (Figure 2). Preterm infants (< 37 completed weeks of gestation) usually have low weight at birth and variable degrees of immaturity in their digestive, respiratory, immune, and neurological systems[60,61], which may require hospitalization for treatments such as intensive antibiotic use, artificial respiration, and parenteral nutrition[62]. Moreover, the common use of C-section for preterm newborns implies the lack of contact with the vaginal/fecal mother's microbiome. Altogether, these situations can lead to an aberrant microbiota composition with increased risks of colonization by pathogenic microorganisms.

Specifically, compared with the age-matched full-term counterpart, colonization trajectories in preterm infants are characterized by a significant reduction in bacterial biodiversity, with low abundance or delayed colonization by the *Bifidobacterium*, and *Bacteroides* genera. Instead, within the first week of life, preterm infants are often colonized by facultative bacteria, such as *Klebsiella*, *Escherichia*, *Streptococcus*, *Enterococcus*, and *Staphylococcus*[63–65] (Figure 2). However, partial normalization of the gut microbiota composition has been demonstrated in preterm newborns receiving breastmilk[66]. These infants exhibited comparable gut microbiomes regardless of birth weight, unlike formula-fed infants whose gut microbiomes still showed distinct clustering patterns based on birth weight.

Postnatal antibiotic treatments can obviously disrupt the normal pattern of colonization and maturation of the infant gut microbiota, inducing long-term

structure alteration[67] (Figure 2). Most studies have agreed that *Bifidobacterium* is the genus most affected by antibiotic use in early life. For example, a recent study observed that, compared with the control group, newborns exposed to antibiotics in the first week after birth experience a diminished abundance of *Bifidobacterium* (*B. adolescentis*, *B. breve* and *B. longum*), *Bacteroides*, and *Escherichia* (*E. coli*) genera, concomitantly with increased abundance of *Klebsiella* (*K. pneumoniae* and *K. oxytoca*) and *Enterococcus* (*E. faecium*) genera[68] (Figure 2). However, reinstatement of *Bifidobacterium* abundance within two to three weeks was favored by short-term antibiotic treatment, while prophylaxis longer than five days results in sustained low levels of these microorganisms for up to six months[69].

In addition to postnatal antibiotic administration, newborns can encounter intrapartum antibiotics either as a prophylactic measure or to address maternal infections. For example, in certain countries, all mothers carrying group B *Streptococcus* receive antibiotic prophylaxis during labor to prevent transmission of this pathogenic bacteria to the forthcoming infant. However, wide-spectrum antibiotics are often used, also limiting the transmission of non-pathogenic bacteria with a similar antibiotic sensitivity profile.

Results from studies examining the effect of intrapartum antibiotics in the gut microbiome structure of healthy vaginally born infants highlighted a marked negative effect on the bifidobacterial community[70–72], with increased abundance of *Staphylococcus* and *Clostridium* genera. In contrast, the impact on members of the *Bacteroides* genus appeared variable, ranging from negative[7] to neutral[70]. Nevertheless, by the age ranging from 12 weeks to 12 months, significant distinctions in the general composition of the infant gut microbiota were no longer observable between those exposed to intrapartum antibiotics and their unexposed counterparts[74]. Hence, it appears reasonable to infer that any potential adverse effects of intrapartum

antibiotic prophylaxis are outweighed by the decrease in early-onset Group B *Streptococcus* infections.

B. From birth to an adult-like gut microbiota structure

Assembly and the main members of the early-life gut microbiota

The prenatal gut can be considered an essentially uninhabited ecological environment which, coinciding with birth, undergoes colonization by a group of initial species denoted as 'founder species', initiating the assembly and the subsequent development of the infant gut microbiota. This formative process follows discernible patterns dictated by the ecological theory of "priority effect", according to which pioneer species anticipate or modify the ecological niche, resulting in the inhibition or facilitation of later colonizing species[75]. Specifically, facultative anaerobic bacteria constitute the initial gut colonizers, falling in the Proteobacteria phylum, including members of *Escherichia*, *Enterobacter*, *Enterococcus* genus derived primarily from the maternal gut microbiome, and Firmicutes, such as *Staphylococcus* and *Streptococcus*, mainly acquired from the maternal skin and oral cavity (Figure 3). These bacteria play a pivotal role in shaping the early gut environment, reducing oxygen levels in the intestine, thereby facilitating the subsequent proliferation of a complex community dominated by obligate anaerobic bacteria, such as members of *Bifidobacterium*, *Bacteroides*, and *Clostridium* (Figure 3).

More specifically, a recent study revealed that the composition of the infant's gut microbiota during the strict lactation period is often dominated by *Bifidobacterium*,

Clostridium, *Streptococcus*, *Escherichia*, and *Klebsiella* genera, which delineated the five main Infant Community State Types (ICSTs)[50]. In a separate study, *Bifidobacterium* genus associated with members of *Enterococcus*, *Lactobacillus*, *Veillonella,* and *Collinsella* genera were recognized as a signature of the 4-month-old gut microbiome[28]. Altogether, these studies concur in supporting that the *Bifidobacterium* genus largely prevails in the early gut microbiota, playing crucial roles in promoting and sustaining the infant's gut development, inhibiting the growth of potentially pathogenic microorganisms, modulating mucosal barrier function, and supporting immunological response maturation[76–79].

Within the infant gut, *Bifidobacterium* abundance begins to increase around the 3rd to 4th day following birth, eventually becoming the predominant genus at approximately one month of age, establishing a highly stable gut-associated community that persists beyond the first year of life (Figure 3).

The ability of bifidobacterial species to successfully colonize and survive in the newborn's intestinal tract is fundamentally linked to their efficient metabolism of the complex oligosaccharides contained in human breast milk (HMOs). Indeed, HMOs serve as a source of nourishment for selected bifidobacterial strains, resulting in a substantial competitive advantage within the infant gut ecosystem. In this regard, a case has been made for the unique genetic arsenal of *B. longum* subsp. *infantis* for degrading a wide variety of HMOs[80]. This species harbors a comprehensive set of enzymes, including fucosidase, sialidase, β-hexosaminidase, and β-galactosidase, and dedicated carbohydrate transporters that collaboratively work to achieve full internal breakdown of complex HMOs.

In contrast, although *B. bifidum* and *B. breve* have shown individual ability to metabolize HMOs only partially, they participate in extracellular breakdown

activities, thus relying on cross-feeding strategies to virtually expand their glycobiome and readily assimilate HMOs[81–83].

From the host perspective, these bifidobacterial species are crucial in promoting and sustaining the infant gut development, inhibiting the growth of potential pathogenic microorganisms, modulating mucosal barrier function, provisioning of vitamins[84], and promoting immunological response maturation[21]. Accordingly, this demonstrates an important microbe-host symbiotic relationship, serving as an example of the intimate (bifido)bacteria-host coevolution underlying the concept of holobiont [85].

The bacterial community established in the infant's gut during the early postnatal period usually remains dominant in the first weeks and months of life, as the microbiome evolves gradually as long as the infant's diet remains uniform. Indeed, while the early gut microbiota during the milk-based diet is enriched in bacteria with genes that facilitate the utilization of milk-derived compounds[86], the subsequent weaning phase, with the introduction of solid foods, gradually promotes the growth of bacteria enriched in protein-encoding genes that enable the utilization of a wider variety of carbohydrates, vitamin biosynthesis, and xenobiotic degradation.

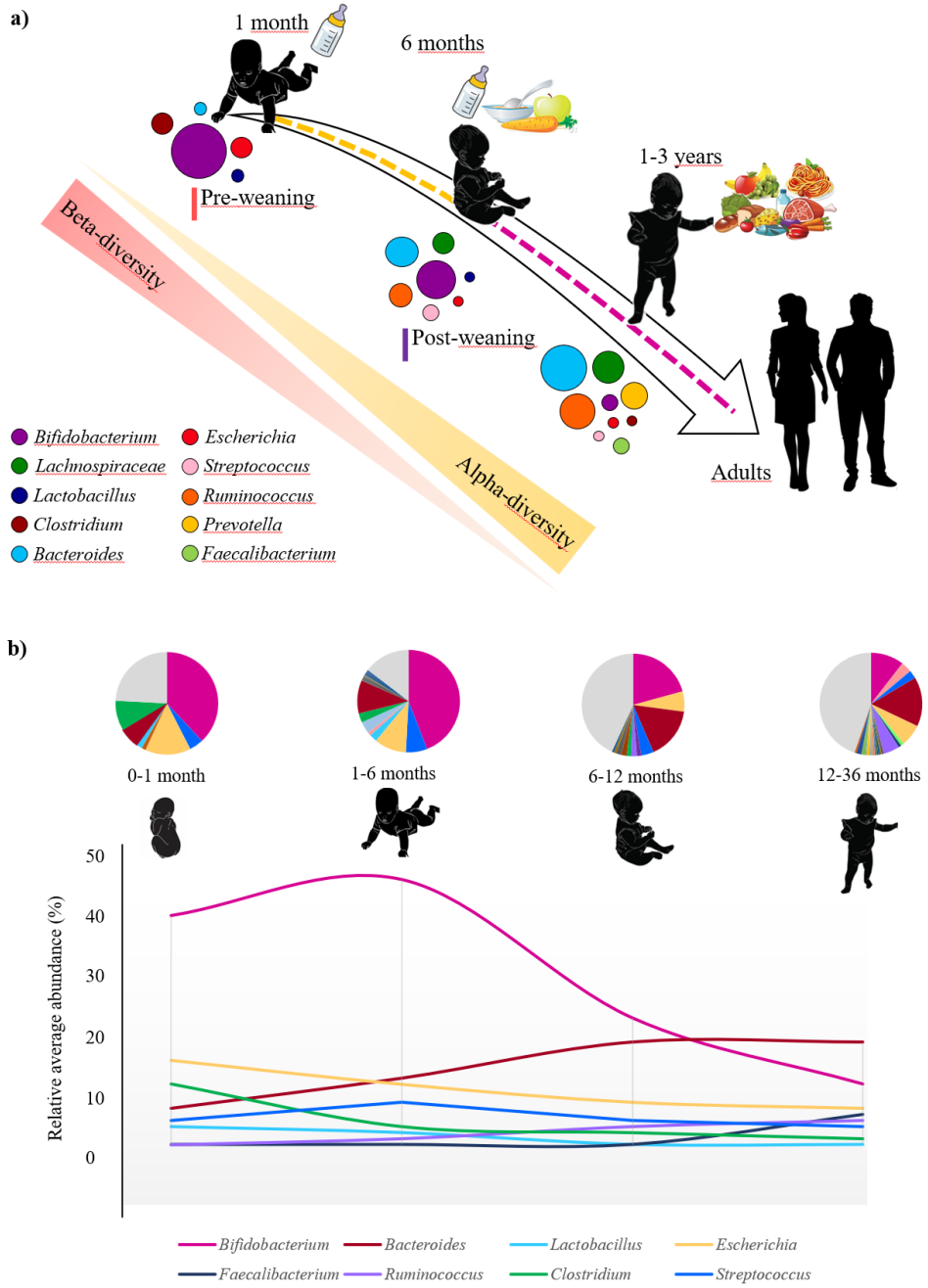


Figure 3. Temporal developmental trajectories of the infant gut microbiota during the first three years after birth. Panel **(a)** shows the role of infant diet in the maturation of infant' gut microbiome structure. Panal **(b)** represents the main age-driven compositional changes of the infant's gut microbial communities.

Maturation of the healthy gut microbiota through infancy

The shared microbiota between mother and child evolves rapidly during the first year of life by acquiring bacterial groups typical of the adult gut microbiota. In particular, infant diet and the weaning period are the most influential factors that shape the early-life microbiota, marking the establishment of a more complex microbiome (Figure 3). Indeed, as the introduction of solid food diversifies the infant's nutritional intake, the nascent gut microbiota adapts to process a wider range of dietary components. This transition is instrumental in shaping the long-term composition and functionality of the microbiome, ultimately influencing the infants' health and metabolism as they grow.

According to The World Health Organization (WHO) guidelines, the optimal time to start providing complementary foods is around the age of 6 months (Figure 3). A large body of studies consistently reported pronounced compositional changes in the healthy infant gut microbiome starting from the weaning, with a steady increase in diversity, decrease in interindividual variations, and shifts in the abundances of major bacterial taxa[28,50,87,88].

As a general rule, the post-weaning age is characterized by a significant reduction in the abundance of *Bifidobacterium* species associated with milk consumption during the suckling phase, being gradually replaced by adult-associated gut commensals, including members of *Ruminococcus*, *Prevotella*, *Eubacterium*, and *Bacteroides* genera[50] (Figure 3). More specifically, according to recent research, at the end of the first year of age, the healthy infant gut microbiota is generally dominated by *Bacteroides* species, accompanied by the emergence of butyrate-producing species such as *Anaerostipes*, *Faecalibacterium*, *Akkermansia*, and *Roseburia*, whose relative abundance tends to increase further as the infants grow older[28,89,90]. This implies a transition towards a more mature intestinal environment correlated with an

enhanced functional capability for carbohydrate degradation. Moreover, at 12 months of age, the infant's gut microbiome is enriched in genes involved in the degradation of complex sugars and starch, as well as vitamins B1, B5, and B12[91].

Interestingly, it has been proposed that, alongside increasing dietary diversity, the cessation of breastfeeding profoundly contributes to shifting the microbial ecology toward an adult-like composition[28,92]. Indeed, continued breastfeeding after solid food introduction determines the persistence of bacterial species associated with an exclusively human milk diet, such as *Bifidobacterium*, *Lactobacillus*, *Veillonella*, and *Collinsella* genera[28]. In contrast, the gut microbiota of 12-month-old infants no longer breastfed was dominated by *Bacteroides*, *Roseburia*, and *Anaerostipes*[28].

Collectively, these investigations suggest a gradual adaptation of the microbiota to utilize the nutrients in the intestinal environment efficiently, depending on breastfeeding status and the introduction of solid foods. While *Bifidobacterium* species occupy key positions in metabolizing HMOs, a variety of bacterial species collaborate in the degradation of proteins and complex plant-derived polysaccharides.

Generally, consensus among most studies indicates that, although the developmental trajectory of the infant gut microbiota is highly individual and dependent on perinatal factors (e.g., delivery mode, feeding regime, gestational age), a large part of maturation in microbial composition and function is completed within the first three years of life, coinciding with the observed stabilization of the infant microbiota into a configuration resembling that of adults[93].

Neonatal intestinal dysbiosis: implication for the infant and long-term diseases

Gut microbiome maturation is intimately linked with the development of the child. Early maladaptive interactions between the nascent gut microbial community and the host, as well as disturbances in the subsequent optimal temporal microbial succession, could impact host's immune functions, intestinal maturation, and neurological development[94–98].

In a general sense, the occurrence of diminished abundance of *Bifidobacteriaceae* (*Bifidobacterium* spp.) alongside high levels of *Enterobacteriaceae* (e.g., *E. coli* and *K. pneumoniae*) and *Clostridiaceae* (*Clostridium* spp.) in the early months of life can be considered as an initial hallmark of neonatal gut dysbiosis. This trend is prevalent in various circumstances, including primarily preterm infants or full-term infants requiring hospitalization or antibiotic administration, and could have immediate health consequences. For example, one of the most severe neonatal diseases recently associated with altered gut microbiota composition is Necrotizing Enterocolitis (NEC), representing a harmful gastrointestinal disease occurring predominantly in premature infants[99]. In NEC events, sections of the infant's gastrointestinal tract experience ischemia and subsequent necrosis, constituting a gastrointestinal emergency in neonates. This condition occurs in approximately 8% of premature infants, with a reported mortality rate reaching up to 25%[100].

Recent studies showed that, compared to healthy premature infants, the immature gut environment of NEC patients is populated by higher levels of *Enterobacteriaceae* genera[101,102]. In this framework, it has been supposed that an intensified pro-inflammatory cascade resulting from a dysfunctional or amplified immunological response to elevated levels of intestinal lipopolysaccharides (LPS) could predispose

infants to NEC pathogenesis[103]. Nevertheless, efforts to identify microbial biomarkers of NEC onset have not yielded consistent results, and thus, further research is necessary to gain insight into the etiology of this neonatal disease.

Beyond NEC, gut dysbiosis can predispose newborns to sepsis by translocation of enteral pathogenic bacteria and dysregulated host response to infection[104]. Preterm and very-low-birth-weight (VLBW) infants are particularly vulnerable to neonatal sepsis, which is typically classified as early-onset sepsis (EOS), appearing within the first 72 h of life, and late-onset sepsis (LOS), occurring more than 72 h after birth [105,106]. While the pathogenic agents of EOS are thought to be obtained through vertical transmission during delivery, LOS pathogens are supposed to be acquired postnatally through the hospital environment[107–109]. Specifically, several studies concluded that decreased diversity and increased abundance of *Staphylococcus* and *Klebsiella* genera and opportunistic Proteobacteria (including *E. coli* and *Pseudomonas* spp) precede LOS onset[110–112]. Moreover, in preterm infants with LOS, bifidobacteria counts were consistently lower[110,113], underlining a disruption in the progression of microbiota from facultative anaerobes to obligate anaerobes dominance.

Inappropriate early-life gut microbiota development can also have long-term adverse health outcomes, [56,93,114–116]. For example, numerous epidemiological studies propose a connection between the early establishment of infant gut microbiota and the risk of allergic diseases in childhood. Specifically, gut microbiota dysbiosis within the first three years after birth with the presence of specific microbial groups has been linked with an increased risk of developing atopic eczema and asthma[115,117]. Indeed, increased levels of *Clostridium* spp., *Escherichia coli,* and *Klebsiella pneumoniae* associated with reduced levels of bifidobacteria were associated with a higher risk of

several atopic outcomes, including eczema, wheezing (asthma), and allergic sensitization[118–120].

The composition and functionality of the gut microbiota can also be implicated in obesity and obesity-related disorders by altering energy harvest. Indeed, gut microbiota composition at two years of age has been linked with Body Mass Index (BMI) at 12 years[121]. Specifically, decreased proportions of *Bifidobacterium* genus concurrently with high levels of *Bacteroides* spp. in early infancy have been linked with a higher risk of pediatric overweight and obesity[122]. Consistently, a recent metagenomic study on the contemporary infant gut microbiome in the USA, known for the high prevalence of childhood obesity, reported a widespread gut dysbiosis with a reduced capacity for microbial HMO utilization.

One of the possible mechanisms underlying the link between the early-life microbiota and metabolic dysregulation appears to be centered in the proportions of bifidobacterial species[123]. Indeed, these latter taxa are the major utilizers of HMOs, whose metabolism generates SCFAs, which contribute to energy homeostasis and modulate host adiposity[124].

Maintenance and restoration of the optimum infant gut microbiome

Although it is not possible to define universally a specific “optimal” gut microbiota composition, the identification of microbial signatures associated with diseased states and the understanding of the risk factors related to early-life dysbiosis have opened novel intervention opportunities. In this context, several studies showed that most aberrant changes in the fecal microbiota of infants could be corrected or mitigated by probiotic supplementation.

Probiotics are living microorganisms known to confer health benefits upon the host, and the supplementation of infant formula milk with probiotics is becoming increasingly common, reflecting an increasing recognition of the potential advantages for the infant's well-being and the establishment of a balanced gut microbiome. Given their well-documented health-promoting effects and importance within the infant gut microbiota, members of the genera *Lactobacillus* (*L. acidophilus*, *L. casei*, *L. plantarum*, and *L. rhamnosus*) and *Bifidobacterium* (*B. bifidum*, and *B. longum* subsp. *infantis*) have been recently used to manipulate the infant's gut microbiota, obtaining promising results in the prevention and treatment of neonatal diseases, including NEC, sepsis, and acute infantile diarrhea[125–129].

Postnatal probiotic administration was also successfully used to prevent or reduce the disruptive effects of antibiotics and Cesarean surgery on the microbiota composition. For example, when associated with breastfeeding, daily oral doses of a multispecies mixture composed of *B. breve*, *L. rhamnosus*, and *Propionibacterium freundenreichii* have been demonstrated to correct the undesired depletion of gut-associated bifidobacteria within cohorts of 3-months-old infants born through Cesarean section and/or receiving antibiotic treatment[130].
Furthermore, in several separate studies, *Lactobacillus reuteri* significantly reduced infantile colic[131,132], while the oral administration of *B. bifidum* has been proved to restore the gut microbiota composition in preterm infants, aligning it more closely with that of full-term infants, thereby reducing the incidence of adverse outcomes[133]. Although the optimal duration of probiotic intervention has not been determined, it could be advantageous to continuously supplement vulnerable infants for the entire critical time of microbiota and immune system maturation.

As discussed above, birth via C-section alters the neonatal microbial profile by preventing the newborn from exposure to the maternal birth canal encompassing the

maternal vaginal and fecal microbiota[46,134]. In this context, an increasing scientific interest has been directed toward assessing the safety and efficacy of postnatal seeding with maternal vaginal or fecal microbiota, which may partially restore the gut microbiota of cesarean-delivered newborns, reducing the risks of C-section-associated diseases[135–137].

Specifically, the temporal development of the gut microbiome composition of the infants seeded with maternal fecal microbiota more closely resembled that of vaginally born infants than C-section infants who were not treated[137]. In contrast, other studies question the value of these practices, observing no modification in the gut microbiome of C-section-delivered infants who received or did not receive maternal vaginal microbes[138]. Overall, further research is needed to confirm the long-term effects, potential benefits, and associated risks of the proposed microbiota restoration therapies.

In the context of infant health, exclusive breastfeeding remains the most optimal and natural way to promote and maintain a healthy neonatal microbiome. In cases where maternal breastfeeding is not feasible, donor breast milk may be an alternative to formula. In the USA, milk banking has been created to collect, process, and distribute human milk donated by nursing mothers who are not biologically related to the recipient infants. Although the pasteurization process to eliminate pathogenic microbes also impacts beneficial bacteria, the polysaccharide component with prebiotic functions, i.e., HMOs, is preserved, promoting the growth of *Bifidobacterium* spp. in the infant gut. However, this resource is typically reserved for premature or fragile infants and can be challenging to access for healthy newborns.

If breast or donor milk is unavailable, fortified formulas containing prebiotic and probiotic compounds can represent actual alternatives. Indeed, recent research has

been aimed at comprehending the intricate microbial interactions in human milk, enabling the optimization of mixtures comprising probiotics (e.g., beneficial bacteria [139]), prebiotics (e.g., HMOs[140,141]), and postbiotic (bacteria metabolites[142]) to emulate the bioactive nourishment provided by breastfeeding[143].

C. Interaction between infant gut microbiota and host's physiology

The influence of the gut microbiome on host metabolism

Most microbial cells inhabiting the human gut are metabolically active, releasing thousands of metabolites at the host-microbiota interface, constituting the gut metabolome. Several of these molecules can act locally, shaping the gut microbiome ecosystem through microbe-microbe interactions and cross-feeding events. However, many others can reach tissues and organs through the blood circulatory system, eventually exerting significant roles in training host immunity, regulating gut endocrine function and neurological signaling, and supplying biological bioactive molecules to the host[144]. The effects of gut microbiota on host's physiology are mediated by metabolites that are either produced by the microbes or derived from the transformation of environmental or host molecules (Figure 4).

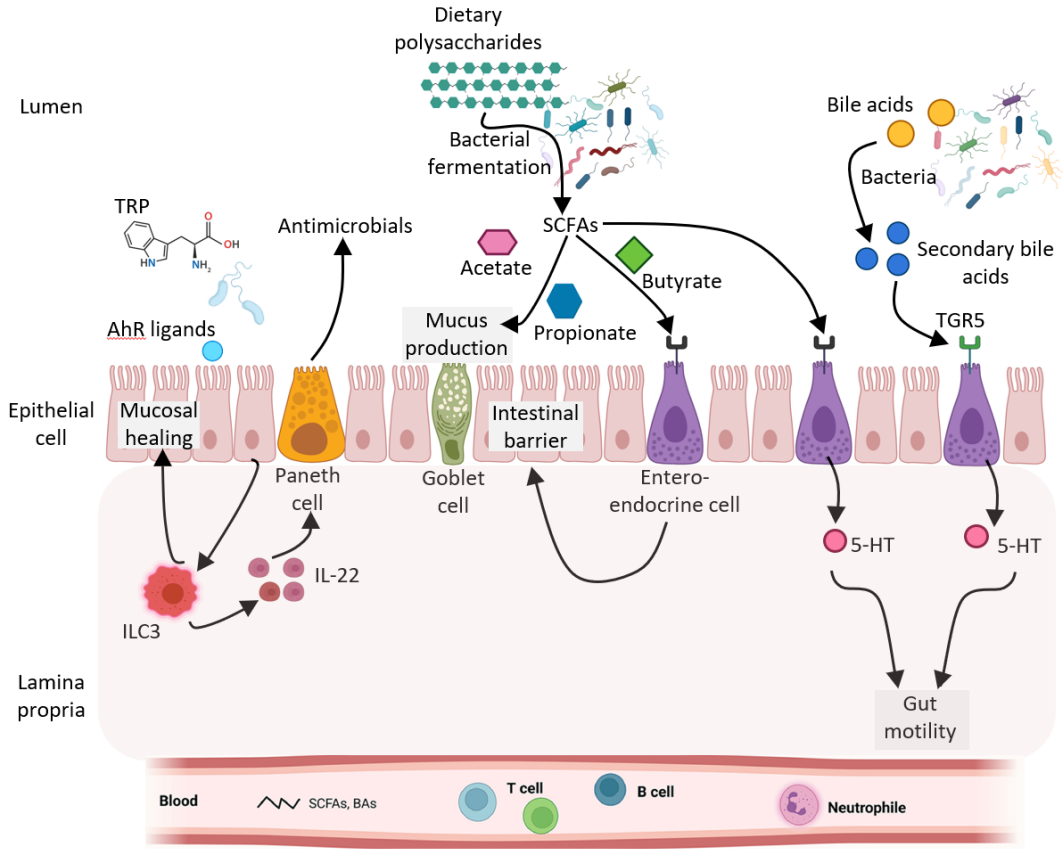


Figure 4. Microbiota and microbiota-derived metabolites effects. Bacteria have the ability to modulate local and systemic environments through direct or indirect pathways. For example, microbial AhR ligands (IAA and IPA) are able to induce mucosal healing and antimicrobial production through ILC3-mediated IL22 secretion. The microbial-derived SCFAs can modulate immune responses and gut motility through the activation of G-protein-coupled receptor. Bile acids can also modulate gut motility by activating the TGR5 receptor on enteric neurons.
Abbreviations: AhR, aryl hydrocarbon receptor; ILC3, innate lymphoid cell 3; TRP, tryptophan; SCFAs, short-chain fatty acids; TGR5, Takeda-G-protein-receptor 5; 5-HT, 5-idrossitriptamina. Created with BioRender.com.

Across infancy, the transition from a milk-based diet to a varied range of solid foods shapes not only the gut microbiota composition as discussed above but also the derived microbial metabolism. For example, during breastfeeding, the dominance of the infant gut by *Bifidobacterium* species degrading HMOs results in the production of short-chain fatty acids (SCFAs) with high proportions of lactate and acetate. In

particular, this latter has been associated with improved intestinal defense against infection and anti-inflammatory responses in the gut epithelium by enhancing the expression of tight junction protein and anti-inflammatory cytokines[145,146] (Figure 4). The bifidobacterial community also converts aromatic amino acids (i.e., tryptophan, phenylalanine, and tyrosine) into aromatic lactic acids (i.e., indolelactic acid, phenyllactic acid, and 4-hydroxyphenyllactic acid), which have demonstrated to exert roles in modulating infant's immune functions[19,147,148]. In accordance with the notion that breastfed infants are dominated by *Bifidobacterium* genus members, aromatic lactic acids have been found with higher levels in feces from breastfed newborns compared to those receiving formula[149,150]or weaned infants[151], emphasizing the interplay between early-life nutrition and specific gut microbes affecting the levels of key microbial metabolites.

With the progression of diet complexity, more indigestible carbohydrates, proteins, and fibers reach the colon, being metabolized by the developing autochthonous microbiome into lactate, pyruvate, and formate during complementary feeding and into propionate and butyrate after cessation of breastfeeding[149,152]. Accordingly, the high abundances of lactate and acetate measured in the infant gut during the early phase reflect a less-developed microbiome since these metabolites typically are converted into butyrate by late infant gut-associated microbial species, such as *Faecalibacterium prausnitzii*, *Roseburia intestinalis*, *Eubacterium rectale*, and *Eubacterium halli* [153]. Consistently, butyrate concentrations are relatively low during milk feeding, steadily increasing across the weaning period and beyond, supporting the growth of intestinal epithelial cells (IECs) (Figure 4). Similar to alterations in microbial composition, deviations in the growth trajectories of gut metabolome may manifest in adverse long-term health outcomes. Indeed, it has been shown that high levels of formate during early infancy (3-4 months of age) have been associated with

a lower BMI z-score in childhood, while high fecal levels of butyrate in early age have been associated with a higher BMI z-score at three years of age[154].

Furthermore, proteins are degraded into amino acids, which are fermented by the resident gut microbes into aromatic acetic and propionic acids (e.g., indoleacetic acid and indole propionic acid), as well as amines (i.e., histamine, dopamine, tyramine, γ-aminobutyric acid [GABA], and tryptamine). In particular, these latter are notorious neurotransmitters that link the gut with the central nervous system (CNS), participating in the gut-brain axis[155,156].

This relationship between the gut microbiota and the brain is pivotal for neurodevelopmental processes and neurophysiology in newborns, contributing to regulating emotions, behavior, and higher cognitive functions. Indeed, autism spectrum disorder (ASD) and attention deficit hyperactivity disorder (ADHD) were found to be associated with changes in the functioning of various neurotransmitters and have also been highly associated with intestinal dysbiosis[157–161].

Among the array of microbe-derived bioactive molecules are also the tryptophan (Trp) metabolites, such as indole and indole-derivatives, including indole-3-acid-acetic (IAA), indole-3-propionic acid (IPA). While IAA and IPA are ligands of the aryl hydrocarbon receptor (AhR), whose activation is considered crucial for intestinal homeostasis by acting on epithelial renewal and barrier integrity[162] (Figure 4), indole is an interspecies signaling molecule that controls aspects of bacterial physiology such as antibiotic resistance, sporulation, and biofilm formation[163,164].

Microbial molecules also can be produced from host-derived metabolites. Gut commensal bacteria, including *Clostridium, Eubacterium, and Bacteroides* species, can convert host-derived primary bile acids into secondary bile acids due to the action of the key enzyme bile salt hydrolase (BSH). Secondary bile acids can then bind the Takeda-G-protein-receptor 5 (TGR5) located in the intestinal endothelial cells[165],

regulating epithelial cell integrity, immune responses, and liver functions[166] (Figure 4). The small intestine during the immediate postnatal period is characterized by the presence of mainly primary bile acids[167,168]. As the infant grows, the intestinal abundance of the microbial BSH gene steadily increases, reaching the highest intestinal levels approximately at one month after birth[168]. Consistently, the hepatic concentration of bacterially modified secondary bile acids increases with infant's age, leading to notable changes in the total bile acid pool after weaning which contribute to regulating lipid, sugar, and energy metabolism in infants[169].

Infant gut microbiota and immune system development

The mammalian immune system is a highly intricate network comprising a complex interplay of innate and adaptive components distributed across various tissues, whose main role is to protect the organism from a multitude of potentially harmful external factors and endogenous imbalances that could disrupt the delicate hemostasis of the body. The first source of antibodies in early life is the passive acquisition from the mother's milk, which contains immunoglobulins A (IgA) and G (IgG). However, the steady increase in gut-associated bacterial diversity necessitates the host to adapt and "learn" how to tolerate the microbiome[170,171]. Indeed, the establishment of the microbial-host symbiosis depends on mutualistic co-development of the host's immune system and the gut microbiota[172].

Multiple studies discuss how the development of the immune system relies on exposure to conserved Microbial-Associated Molecular Patterns (MAMPs)[173,174]. For example, early responses to microbial ligands such as lipopolysaccharide (LPS), the endotoxin located in the outer membrane of Gram-negative bacterial walls, train gut epithelial cells to be less responsive to subsequent TLR stimulation, generating

protective innate immune tolerance[175–177] (Figure 5). Similarly, recent studies have revealed that *Bifidobacterium longum* subsp. *infantis* can stimulate dendritic cells (DCs) to induce proliferation of regulatory T (Treg) cells, which play a central role in the regulation of immune responses against self-antigens, allergens, and commensal microbiota (Figure 5). Activated DCs may also induce helper T (Th) cells responses and secrete anti-inflammatory cytokines (Figure 5), thus promoting intestinal mucosal homeostasis[178,179]. Similarly, Treg cells are demonstrated to be induced in the colon also by *Bacteroides fragilis* via polysaccharide A (PSA)[180], as well as by many other gut commensals[181,182], highlighting the importance of this pathway for the control of mucosal-microbiota homeostasis via immunological tolerance.

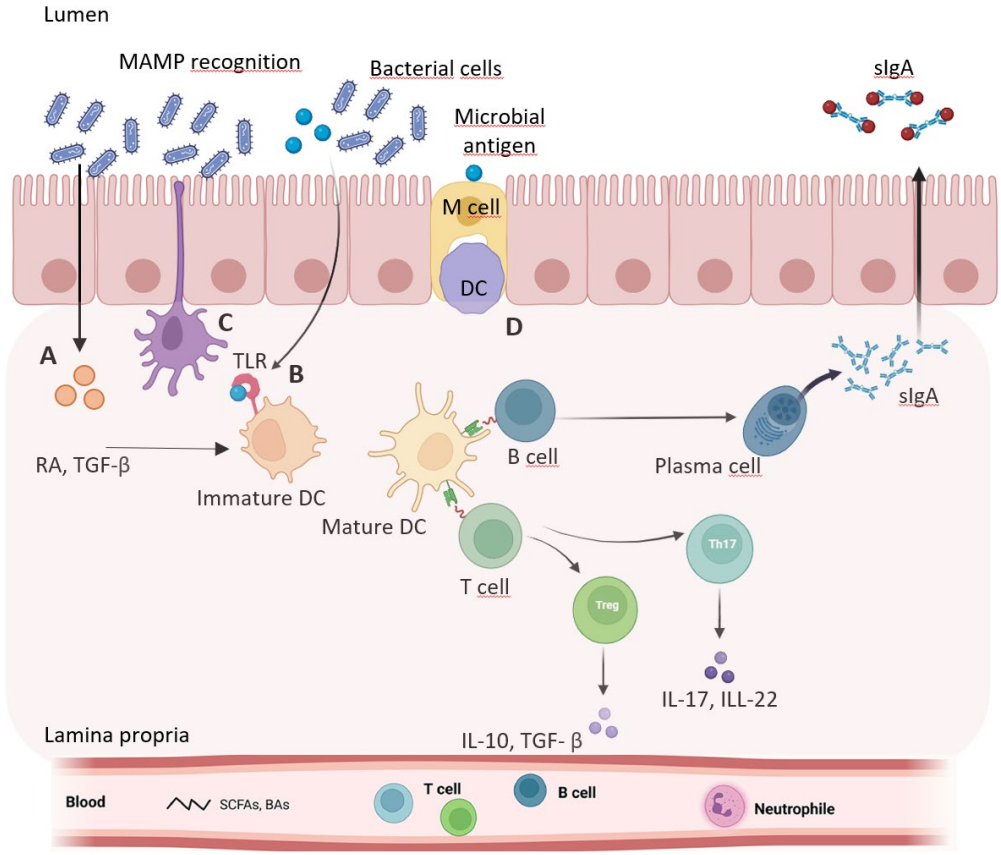


Figure 5. Gut microbiota and host immunity. Microbe-associated molecular patterns (MAMPs), expressed on the bacterial surface, are recognized by pattern recognition receptors (PRRs), expressed by intestinal epithelial cells (IECs). Signals from the gut microbiota is transferred by IEC-released factors, such as retinoic acid (RA) and TGF-β **(A)**, or throught the direct binding to TLRs **(B)**, eventually promoting the development of mature DCs that stimulate the differentiation of T cells into Treg and Th17. Th17 cells can produce IL-17 and IL-22, promoting inflammation at the local site and recruiting neutrophils from the blood. In contrast, Tregs secret IL-10 and TGF-β, which can inhibit the activity of multiple immune cells and create an anti-inflammatory cytokine milieu. Luminal antigens can also be captured directly by DCs through extension of dendrites (membrane extensions) between epithelial cells **(C)**, or indirectly by endocytosis of specialized antigen-sampling cells (called M cells) and transcytosis to DCs **(D)**. B cells differentiate into plasma cells (PCs) secreting IgA that translocate through the epithelium and are released into the mucus layer where control bacteria adhesion to host tissues.
Abbreviations: PGN, peptidoglycan; LPS, lipopolysaccharides; DCs, dendritic cells; sIgA, secretory IgA; MAMPs, microbial-associated molecular patterns; TLR, Tall-like receptor; M cells, microfold cells. Created with BioRender.com.

However, rather than specific bacterial taxa, most studies point to global changes in the microbial community (diversity and metabolite shifts) as drivers of immune development[183]. In this context, increasing evidence showed that the weaning period is critically important in the imprinting of the immune system through the so called "weaning reaction"[184]. Indeed, it has been demonstrated that increased bacterial richness during the weaning period leads to a vigorous immune reaction characterized by a transient production of pro-inflammatory tumor necrosis factor alpha (TNFα) and interferon gamma (IFNγ) by lymphocytes (T cells)[184,185]. Interfering with this process results in an inappropriate imprinting of the immune system and subsequent increased susceptibility to allergy, colitis, and cancer later in life. Indeed, microbial colonization after the weaning period cannot compensate for the lack of microbiota-induced immune stimulus in early life and the weaning reaction, providing further evidence that the maturation of the immune system in infants relies on a temporally structured succession of the gut microbiome.

Early-life immune development is also reliant on the actions of a group of bacterial metabolites known as short-chain fatty acids (SCFAs), which are direct byproducts of bacterial colonic metabolism.

SCFAs, and in particular butyrate, are essential energy sources for colonocytes[186], and have been shown to induce proliferation and differentiation in the gut environment of T cells, such as Treg and Th cells[187,188], and B cells, such as IgA or IgG-secreting B cells[189].

Collectively, compelling evidence suggested that key taxa, microbial community, and bacterial metabolites represent modulatory triggers of host immune function maturation by influencing the repertoire, amount, and activation of the cellular component of the immune system. This contributes not only to the maintenance of

the immune system-microbiota alliance but also to the systemic regulation of immune responses.

D. Long-lasting effect on the gut microbiota.

Impact lifestyle and diet on the gut microbiota

As mentioned above, the infant gut microbiota is believed to achieve full maturation and an adult-like conformation within the third year after birth. However, although whole shotgun sequencing data show stability over time[190], the gut-associated microbial community continues to act as a transducer of environmental signals lifelong, undergoing compositional and functional shifts in response to varying circumstances. In this context, dietary and non-dietary lifestyle habits can shape the gut microbiota in favor of altered composition that increases the risk of adverse health outcomes (Figure 6).

For example, recurrent intake of a wide range of common non-antibiotic drugs has been associated with specific microbial signatures[191]. Among these, proton-pump inhibitors (PPIs) appear to have large effects on the gut microbiome composition, with an increased abundance of *Streptococcus mutans* and *Veillonella parvula* known to typically inhabit the oral niche[192]. While such effect of PPIs use can potentially be explained by changes in acidity that facilitate the growth of upper intestinal bacteria in the gut, a direct inhibitory effect of PPIs was observed for *Dorea* and *Ruminococcus* species[193–195].

Psychological stress is another lifestyle factor that can affect the activity of the colon through the bidirectional gut-brain axis[196]. Psychological stressors manifest in various forms, impacting individuals during both early childhood and adulthood, with outcomes spanning from acute and chronic negative effects on health. Recently,

acute, and chronic stress has been associated with altered gut microbiota profiles. Specifically, although alpha- and beta-diversity did not appear significantly affected[197–199], the relative abundance of *Lachnospira*, *Veillonella*, and *Sutterella* decreased in high stress, while *Roseburia* and *Rhodococcus* were increased[199]. In another study, stressful life events were consistently associated with the reduced abundance of *Bacteroides* genus in children[199].

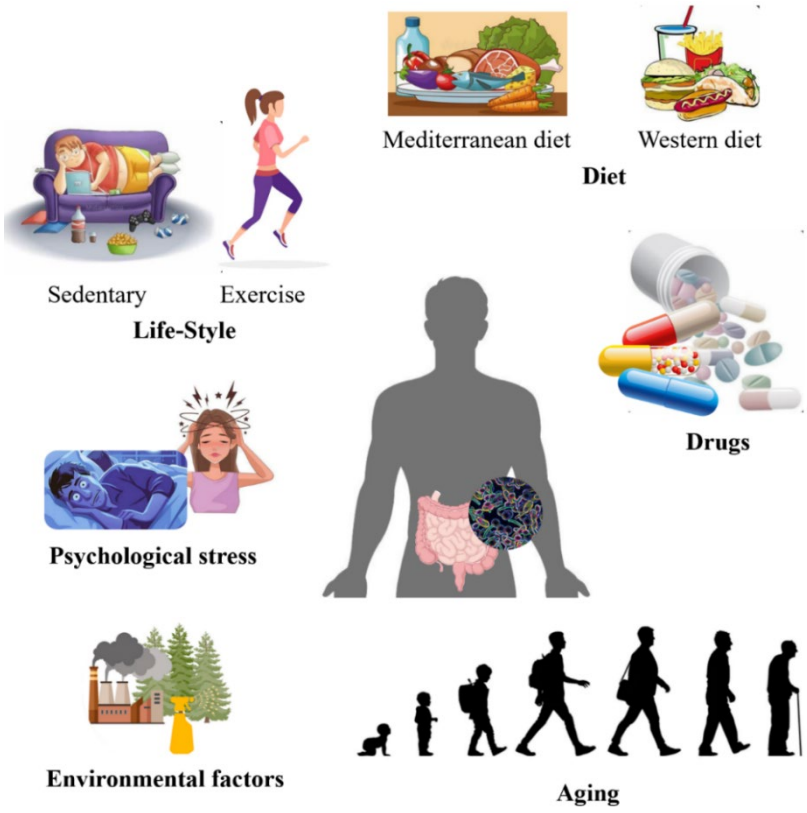


Figure 6. Factors influencing the human gut microbiome throughout lifespan.

Despite the observed connections between psychological stress and gut microbiota composition, the underlying mechanisms remain unclear. However, emerging evidence proposes that dysregulation in tryptophan metabolism and specific

hormones such as catecholamines and cortisol might contribute to these associations[200,201].

Furthermore, sedentary behavior and a lack of physical exercise are known to affect the composition, diversity, and function of the gut microbiota regardless of age and diet[202–204] (Figure 6). Indeed, gut microbial diversity is decreased in a sedentary lifestyle, leading to an increasing incidence of chronic diseases[205]. In contrast, physical activity promotes the growth of bacteria that modulate mucosal immunity, enhance barrier functions, and produce bioactive metabolites, such as short-chain fatty acids (SCFAs), providing protection against gastrointestinal disorders, colon cancer, obesity, and metabolic diseases[206,207].
However, the mechanism by which physical activity induces these changes is not fully understood. Likely, it involves various interconnected factors and pathways, including alterations in the bile acids profile with potential antimicrobial effects, increased production of IgA linked to resistance against certain microorganisms, elevated production of SCFAs, suppression of Toll-like receptor signaling pathways leading to reduced serum LPS levels, and reduction of intestinal transit time[208].

Finally, the source, quality, and type of food clearly influences the composition of the microbiota and host-microbe interactions (Figure 6). For example, animal studies report that while monounsaturated and polyunsaturated fats can have beneficial effects on the host microbial ecosystem and anti-inflammatory response, dietary saturated fats have detrimental effects on the gut microbiota, intestinal permeability, and inflammatory status[209,210]. This latter is notably evident in Western lifestyle, which is associated with obesity and related comorbidities, including diabetes type II, metabolic syndrome, and heart disease[211], as well as chronic diseases and intestinal inflammation[212]. Indeed, Westernized diet is characterized by a combination of high

fat/high sugar diet, which creates a specific inflammatory environment in the gut, thereby increasing host susceptibility to chronic inflammatory bowel disease[212,213]. Moreover, it has been proposed that the low-fiber intake of a Western diet leads to the depletion of *Akkermansia muciniphila* and various butyrate-producing bacteria, resulting in the absence of protective functions against inflammation and barrier permeability[213]. In contrast, the high intake of fiber and complex carbohydrates characterizing unindustrialized/rural population, as well as Mediterranean or vegetarian diet, are related to higher abundance and prevalence of *Ruminococcus*, *Faecalibacterium, Eubacterium, Akkermania,* and *Prevotella* genera[214–217], linked with vitamins and SCFAs production, reduced risks of gastrointestinal inflammation status, as well as cardiovascular and metabolic diseases.

The gut microbiota in aging

As described above, the human gut microbiota is established after birth, starting as a dynamic ecosystem that stabilizes within the first 2-3 years. Subsequently, the structure of this gut-associated microbial community increases in both diversity and richness, reaching a homeostatic climax and its highest complexity in adulthood. Nevertheless, at the later stages of life, the gut microbiota composition becomes again less diverse and more dynamic[217,218], characterized by a higher abundance of Proteobacteria and Bacilli, with lower levels of *F. prausnitzii*[218–220]. Moreover, the age-related decrease in immune system functions, known as immunosenescence 19047800, is characterized by a persistent low-grade inflammatory state, termed inflammageing 18240544, 18400689, which contributes to undermining the homeostatic equilibrium in the gut, leading to changes in intestinal microbiota structure and activity.

Not surprisingly, advancing old age also affects the production of bacterial metabolites. For example, SCFAs are produced with a lower concentration in elderly, likely because of variations in the colonic bacterial metabolism that shift from the predominantly saccharolytic metabolism observed in adults toward a predominantly putrefactive metabolism 15466557. Consistently, several butyrate-producing bacteria, including *Ruminococcus obeum*, *R. intestinalis*, *E. rectale*, and *E. halii*, were found with low abundance in older individuals.

Interestingly, although 65 years is commonly utilized to classify physiological senescence according to WHO standards, a more recent study sampling densely individuals between 65 and 100 reports that the impact of the aging process on the gut microbiota becomes noticeable around the ages of 75-80 years[221]. Such changes in the composition and functionality of the intestinal microbiota in older adults coincide with modifications in the physiology and functions of the digestive tract, as well as clinical conditions and diseases, such as *Clostridium difficile* colitis, colon cancer, cardiovascular disease, and systemic inflammation. Discerning whether the observed alterations in gut microbiota composition in age-associated diseases are causal factors or mere consequences remains challenging.

However, several studies propose that manipulating the microbiota in the elderly population is a promising strategy to prevent or reduce aging-associated diseases and conditions[222–227]. From these works, it emerges that supplementations with strains of the *Lactobacillus* and *Bifidobacterium* genera can effectively improve digestive and immune functions, as well as alleviate various aging-related disorders, including constipation, diarrhea, and bone loss.

Nevertheless, the effect of probiotic supplementation targeting fundamental aging processes in humans remains inconclusive, and further investigations are still needed

to gain insight into the mechanisms by which probiotics can potentially impact healthy aging.

Chapter 2

Outline of the thesis

The purpose of this Ph.D. thesis is to provide insights into the dynamic changes underlying the assembly and maturation of the infant gut microbiota. It is well known that the first three years after birth represent a crucial moment for the establishment of the human gut-associated microbial community, being influenced by perinatal environmental conditions and host factors, including primarily the mode of delivery, feeding practices, use of antibiotics, gestational age, and the introduction of solid foods. These early-life circumstances are known to have profound and lasting implications for an individual's health, affecting short- and long-term susceptibility to various diseases. Therefore, understanding the mechanisms fundamental to the establishment of human gut microbiota is crucial for promoting its optimal development and, ultimately, lifelong well-being.

Chapter 3 investigates the strain-level dynamics of infant gut microbiota members through infancy, as well as the host-microbe interactions underlying the persistence of early key microbes throughout the lifespan. In this context, multi-omics approaches allowed us to highlight a longer-lasting colonization of *B. longum* and *B. bifidum* preferentially in the intestine of women compared to males, likely as microbe reservoirs for future generations.

Chapter 4 explores the genomic diversity of the *B. longum* taxon through comparative genomic and phylogenomic approaches, highlighting a remarkable genetic and phenotypic diversity within this species. Additionally, this study examines horizontal gene transfer events, suggesting a role in conferring this species with genetic features for increased competitiveness in the gut environment of suckling infants.

Chapter 5 describes the variation at the single nucleotide level within protein-encoding genes shared across human-derived *L. crispatus* strains. This taxon is known to inhabit the human vaginal niche, contributing to establishing e healthy environment. Findings of this work highlighted varying growth performances among

members of this species, suggesting that the colonization and stable persistence within the female reproductive tract could be strain-dependent.

Chapter 6 discusses the phylogenetic organization of the *Gardnerella vaginalis* taxon, proposing the existence of nine genotypes associated with different virulence potentials, suggesting the presence of both putative pathogenic and commensal *G. vaginalis* strains. These insights, coupled with metagenomic microbial profiling of human vaginal microbiomes, provided an understanding of genotype distribution across the population, underscoring the existence of genetically diverse *G. vaginalis* communities, which may impact the risk of bacterial vaginosis.

Chapter 7 inspects the gut microbiota composition of infants affected by NEC and before NEC diagnosis, identifying members of the *Clostridium* genus and lactate as possible early biofunctional markers of NEC onset.

Chapter 8 considers common bifidobacterial strains used in probiotic products to build a genomic database named Integrated Probiotic DataBase (IPDB). The IPDB was then utilized to perform comparative genomics analyses of genetic features conferring structural, functional, and chemical characteristics possibly involved in microbe-host and microbe-microbe interactions. Accordingly, the IPDB represents a useful tool to rapidly access the genetic potential of widely-used bifidobacteria probiotic strains, providing precise evidence behind the claimed beneficial effects of each probiotic.

Chapter 9 details the dynamic changes in the microbial metabolism related to the production of bioactive small molecules during the maturation stages of the human gut microbiome. Moreover, metagenomic data from breastfed and formula-fed infants were also investigated to inspect how feeding types can modulate the metabolic functionality of the early gut microbiome.

Chapter 10 aims to explore the long-lasting bidirectional relationship between the gut microbiome and its host by performing a species-level meta-analysis. This survey considers host' lifestyles as a factor that profoundly modulates the microbiome structure and functionality, revealing significant differences between sedentary and athletic individuals.

Chapter 3

Genetic strategies for sex-biased persistence of gut microbes across human life

Chiara Tarracchini, Giulia Alessandri, Federico Fontana, Sonia Mirjam Rizzo, Gabriele Andrea Lugli, Massimiliano Giovanni Bianchi, Leonardo Mancabelli, Giulia Longhi, Chiara Argentini, Laura Maria Vergna, Rosaria Anzalone, Alice Viappiani, Francesca Turroni, Giuseppe Taurino, Martina Chiu, Silvia Arboleya, Miguel Gueimonde, Ovidio Bussolati, Douwe van Sinderen, Christian Milani*, Marco Ventura*.

The results of this chapter were published in Nature Communication, 2023 Jul; doi: 10.1038/s41467-023-39931-2.

*These authors contributed equally.

Abstract

Although compositional variation in the gut microbiome during human development has been extensively investigated, strain-resolved dynamic changes remain to be fully uncovered. In the current study, shotgun metagenomic sequencing data of 12,415 fecal microbiomes from healthy individuals are employed for strain-level tracking of gut microbiota members to elucidate the evolving biodiversity across the human lifespan. This detailed longitudinal meta-analysis reveals host sex-related persistence of strains belonging to common, maternally-inherited species, such as *Bifidobacterium bifidum* and *Bifidobacterium longum* subsp. *longum*. Comparative genome analyses, coupled with experiments including intimate interaction between microbes and human intestinal cells, show that specific bacterial glycosyl hydrolases related to host-glycan metabolism may contribute to more efficient colonization in females compared to males. These findings point to an intriguing ancient sex-specific host-microbe coevolution driving the selective persistence in women of key microbial taxa that may be vertically passed on to the next generation.

Here, via analyses of shotgun metagenomic sequencing data of more than 12,000 fecal microbiomes from healthy individuals, the authors reveal the presence of microbiome genetic traits involved in host mucin metabolism, supporting colonization and persistence of specific bacterial strains preferentially in the intestinal environment of women compared to men.

For Supplementary Materials see the article published in Nature Communication

Introduction

The intestinal microbial community of an infant gradually assembles through a patterned developmental process following birth. In particular, the first three years of life represent a critical window of opportunity during which short-term changes in the gut microbiota composition occur in conjunction with rapid physical development of the newborn[1,2]. This initial dynamic process eventually evolves towards stable microbe-host interactions that are of paramount importance to impart beneficial effects on host health, such as metabolism of non-digestible dietary carbohydrates and stimulation of endogenous intestinal mucus production, vitamin synthesis, development and homeostasis of the immune system, as well as protection against pathogens[3]. In particular, it is known that, at species level, the establishment of a stable gut microbiota occurs through two main diet-guided stages in early childhood[4,5], with the first one taking place immediately after birth when exclusive milk-feeding begins[6,7]. This stage is characterized by a gut microbial community which, upon being partly (vertically) transmitted from the mother and partly acquired through contact with the surrounding environment during and after delivery[8–10], represent the first microbial colonizers of the infant gut due to their ability to directly or indirectly metabolize human milk oligosaccharides (HMOs)[11,12]. Another transition occurs during the weaning period, typically around the age of six months, when infants are gradually introduced to a solid and more varied diet, which offers new ecological niche colonization opportunities[13–15].

Whole-metagenome shotgun (WMGS) sequencing represents a powerful tool to disentangle the composition of complex microbial communities and retrieve genomes belonging to hard-to-culture microbial species[16–18]. Accordingly, several recent longitudinal studies have investigated the infant gut microbiome composition, highlighting the sequential age-related changes at species level[19,20]. For example, bifidobacterial species, key microbial taxa of the infant gut microbiota, can persist at a lower level (2–14% relative abundance) throughout adulthood and can

subsequently be passed on to the next host generation by mother-to-infant vertical transmission[21,22]. Nevertheless, the impact of the host biological sex on the assembly and maintenance of the gut microbial community remains poorly investigated.

In the current study, a total of 400 longitudinal fecal metagenomes from 124 healthy infants (0-3 years) were investigated to inspect the intra-species variations underlying the assembly of the infant gut microbiome during the first two years following birth. In this context, we assessed if particular gut microbiota members elicited higher persistence in female infant (compared to male counterparts), perhaps to maintain cross-generational transmission. These analyses, coupled with inspection of shotgun metagenomic data from 12,415 healthy subjects (6545 females and 5870 males) aged from a few days to 90 years, allowed the identification of two glycosyl hydrolases, i.e., members of GH101 and GH136, which appear associated with persistence of *B. bifidum* and *B. longum* strains with preferential presence in the female gastrointestinal tract. Moreover, sex-related (bifido)bacterial resilience was validated *in vivo* by retrospective human clinical trial data involving the supplementation of bifidobacterial strains displaying a persistent vs. non-persistent genotype.

RESULTS

Strain dynamics of the gut-associated microbiota within the first 24 months of life

Shotgun metagenomics sequencing approaches were applied to the microbiomes of 11 healthy, vaginally delivered, full-term (>37 weeks of gestation) newborns, which were longitudinally sampled at 1-, 6-, 12-, and 24-months following birth (Figure S1, Figure S2, Table S1). Consistent with previous scientific literature focusing on infant community state types (ICSTs)[24], the species-level taxonomic classification of the sequenced reads revealed that *Bifidobacterium longum* and *Escherichia coli* were the most prevalent microbial infant gut components at the pre-weaning stage, followed by *Bifidobacterium pseudocatenulatum, Bifidobacterium breve, Collinsella aerofaciens*, and *Bifidobacterium bifidum* (Table S2). Furthermore, as reported previously, the relative average abundances of these species decreased in post-weaning (Kruskal-Wallis test, p-value < 0.01), simultaneously with the progressive colonization by adult-associated bacterial species, including *Eubacterium rectale*, *Faecalibacterium prausnitzii*, *Ruthenibacterium lactatiformans*, *Akkermansia muciniphila*, and members of the *Bacteroides* genus, such as *Bacteroides uniformis* (Figure S2, Table S2)[23,24].

With the aim of investigating strain dynamics in the developing infant gut ecosystem, a total of 63 metagenomically assembled genomes (MAGs), corresponding to the 11 above-listed main gut-associated microbial species (Table S3), were coupled with conspecific publicly available genome sequences in order to build 11 species-specific databases of reference strains (Table S4). Following assessment for completeness (>90 %) and ANI-driven dereplication, these 11 databases were employed to

investigate strain-specific persistence and stability, i.e., the time span during which longitudinal samples harbored identical strains (Figure S3)[25].

Collected data revealed that strains belonging to species associated with the introduction of solid diets, such as *E. rectale* and *C. aerofaciens*, assemble into heterogeneous strain communities, appearing vulnerable to intestinal niche changes that occur during the first two years of infant life (Figure S3). Conversely, specific *B. longum* subsp. *longum*, *B. bifidum, B. breve*, and *B. pseudocatenulatum* strains established stable host-microbe symbiotic relationships lasting beyond the weaning phase in 91%, 72 %, 54.5 %, and 45.4% of the inspected infants, respectively (Table S2, Figure S3), thereby representing the most resilient, unvarying, and stable bacterial communities of the assessed infant gut microbiome (See supplementary text for details). Specifically, after accounting for sequencing depth, an adjusted average number of 1.58 *B. longum* subsp. *longum*, 1.44 *B. bifidum*, 0.82 *B. breve*, and 0.80 *B. pseudocatenulatum* strains were found to be shared among multiple time points within the first birth year (Figure S3), thusn emerging as significantly more persistent than other principal members of the suckling infant's gut microbiome (Kruskal–Wallis with Dunn's post-hoc test, *p*-values after the Benjamini-Hochberg correction < 0.05; Table S5; Figure S4). Moreover, specific *B. bifidum* and *B. longum* subsp. *longum* strains were detected in the infant gut up to the second year after birth, coinciding with the last sampling time (Table S5). Consistently, these bifidobacterial species are known to be genetically adapted to colonizing the infant intestine, due to particular metabolic activities and cooperative trophic interactions, i.e., cross-feeding actions, toward complex carbohydrates, such as HMOs and human mucin[26–28].

The strain-resolved dynamics of the decoded infant gut microbiome were experimentally validated using strain-specific primers through quantitative real-time PCR (qRT-PCR) (Table S6), confirming that fluctuating and/or persistent patterns

occurring during the infant gut microbiome development are strictly dependent on the microbial species.

To validate the results from our population-wide metagenomic study and to ensure that the bioinformatic approaches did not bias the observed bacterial persistence patterns, an independent validation cohort was constructed employing a large, publicly available infant dataset that includes fecal samples from multiple time points spanning the first year after birth[29] (Table S1). The strain-resolved microbial community composition of this particular metagenomic dataset was determined employing the pipeline implemented by Mäklin et al.[30] As detailed in Table S7, the development trajectories of the infant gut microbiome observed within the validation cohort confirmed what we noted in our study population, highlighting a longer-term persistence of bifidobacterial strains compared to those belonging to non-bifidobacterial species, i.e., *C. aerofaciens*, and *E. coli* (Chi-Squared post-hoc test, p-value < 0.05) (Table S7). Moreover, a statistically significant higher persistence of the early *B. bifidum* and *B. longum* colonizing strains was observed in vaginally delivered infants when compared to those born by C-section (Chi Squared test, p-values < 0.05) (Table S7). These results expand on our previous findings, suggesting a potential greater ability of maternally-derived bifidobacterial strains to persist in the infant gut microbiome during the neonatal period.

Correlation between microbial resilience and host sex

As we mentioned above, the first seeding of (members of) the gut microbiota is believed to occur during delivery, involving the transfer of microbial lineages from the mother to the respective newborn [22,31]. In this context, we questioned if microbial strain persistence is more effective in females to sustain the intergenerational transmission of ecologically well-adapted gut microbiota members. For this purpose,

the vaginally delivered fraction of the above mentioned longitudinal metagenome infant dataset (72 females and 73 males, for a total of 145 vaginally delivered infants) were integrated with 357 additional, publicly available shotgun samples from longitudinal studies of the infant gut microbiome (54 females and 59 males), encompassing pre-weaning (0-6 months) and post-weaning (over six months) time-points (Figure S1, Table S1)[11,32]. All infants were healthy, delivered vaginally at term, and were not subjected to antibiotic treatment (Table S1). Hence, considering the above-described high persistence of bifidobacterial strains throughout the transition phase of weaning, we evaluated the sex-specific stability of strain communities belonging to *B. bifidum*, *B. longum* subsp. *longum*, *B. breve*, and *B. pseudocatenulatum* between pre- and post-weaning stages by using the StrainGE tool (see Methods).

Specifically, for each infant, the bifidobacterial strain dominating the gut microbiome at 0-6 months of age was compared with those preeminent after the introduction of complementary solid foods (around 12 months). At species level, *B. bifidum* was detected from lactation to the post-weaning phase in 31 % of the 258 inspected infants, including 36 females and 44 males (Figure 1a, Table S9). While no sex-related difference in species-level stability was observed across weaning (Fisher test, *p*-value=0.349), the dominant *B. bifidum* reference strains detected at 0-6 months persisted after the weaning phase in 24 out of 36 female infants (67 %) and in 15 of the 44 males (34 %) (Figure 1a, Table S9), suggesting greater strain-level stability in the large intestine of female infants (Fisher test, *p*-value=0.007).

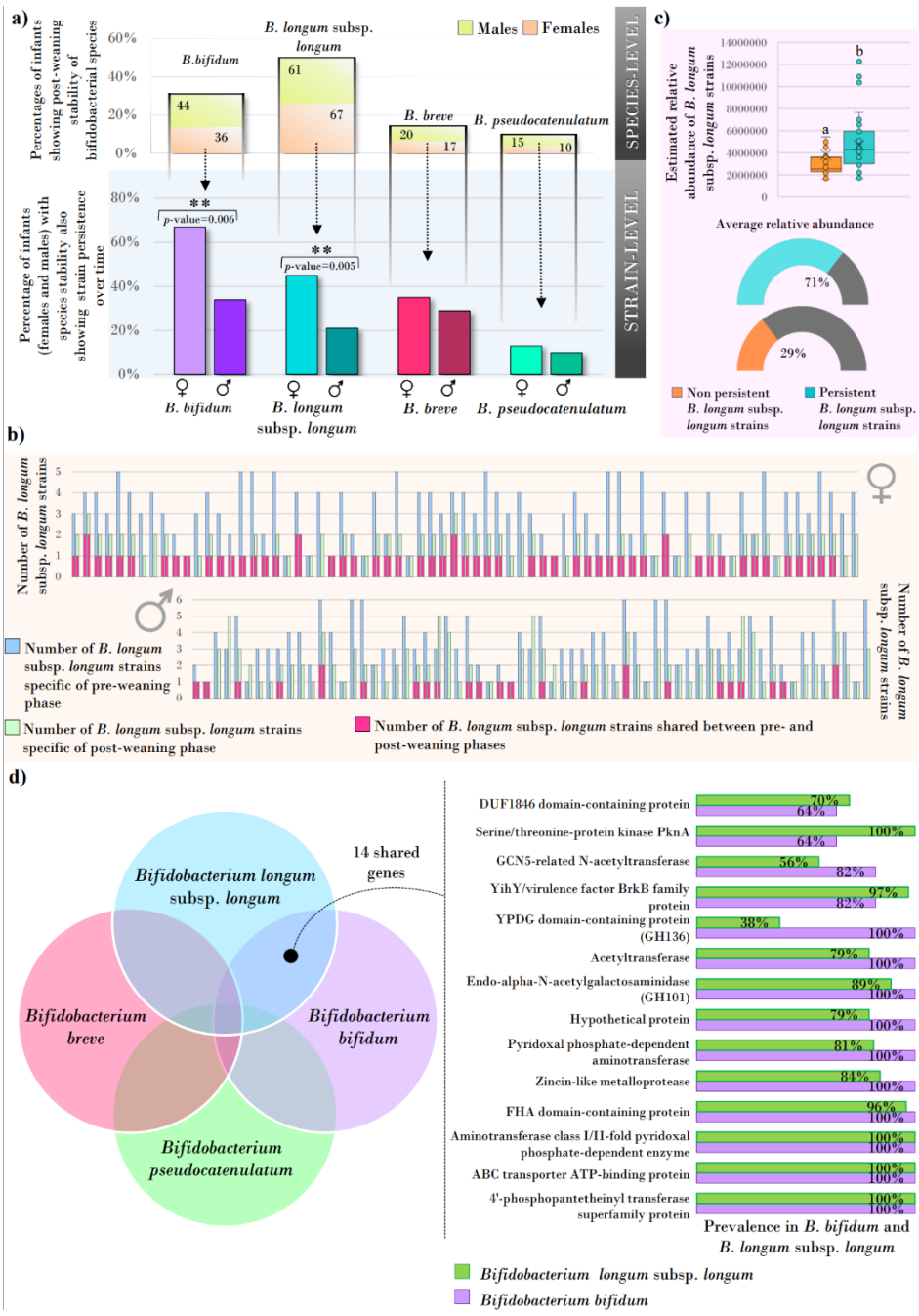


Figure 1. Gender-specific persistence of *B. bifidum*, *B. longum* subsp. *longum*, *B. breve*, and *B. pseudocatenulatum* species and the 14 genes shared between the female-associated persistent bifidobacterial species. In panel **(a)**, the bar chart on the top displays the species-level persistence from pre- (1–6 months) to post-weaning stages (12–24 months) of *B. bifidum*, *B. longum* subsp. *longum*, *B. breve*, and *B. pseudocatenulatum* in the infant population. Below, bar plots represent the 12–24-months gender-specific stability of the bifidobacterial strains, expressed as the percentage of (infant) females and males showing

after-weaning persistence of the same bifidobacterial strains identified at 0–6 months. Statistically significant gender-related differences were highlighted with an asterisk on the top of the columns (Fisher test, *p*-value = 0.006 and p-value = 0.005). Panel **(b)** refers to the inspection of the whole *B. longum* subsp. *longum* strain communities in female and male infants. Each pair of bar plots shows the number of strains identified in the pre-weaning (left) and post-weaning phase (right). Different colors highlight the number of *B. longum* subsp. *longum* strains found only in pre-weaning age (light blue), only in the post-weaning phase (light green), and shared between pre- and post-weaning time-point (pink). Panel **(c)** depicts the statistically significant difference in an estimated relative abundance (obtained by normalizing the genome coverage on the corresponding genome length) of persistent (n = 83, light blue) and non-persistent (n = 171, orange) *B. longum* subsp. *longum* strains in the infant post-weaning time points (Mann–Whitney test, *p*-value = 0.001). The boxes are determined by the 25th and 75th percentiles. The whiskers are determined by 1.5 interquartile range (IQR). The line in the boxes represents the median, while the cross marker (X) represents the average. In panel **(d)**, the Venn diagram on the left highlights the 14 genes shared by *B. longum* subsp. *longum* and *B. bifidum*, while bar charts on the right-side report the prevalence of each PDC in publicly available complete genomes *B. longum* and *B. bifidum*.

Similarly, 49.6 % of all assessed infants (67 females and 61 males) exhibited cross-weaning persistence of the *B. longum* subsp. *longum* species, whose dominant reference strains identified at 0-6 months were also found at 12-24 months in 45 % of the females, being significantly higher than what was observed for male infants (21 %) (Fisher test, *p*-value = 0.005) (Figure 1a, Table S9). Specifically, it appears that males undergo higher fluctuations in (bifido)bacterial strain composition compared to female infants, whose early-engrafted dominant *B. bifidum* and *B. longum* subsp. *longum* strain was maintained through the weaning phase with a higher frequency.

In contrast, *B. breve* and *B. pseudocatenulatum* species were detectable across time-points (pre- and post-weaning) only in 14.3 % and 9.6 % of infants, respectively (Figure 1a), showing no significant sex-associated difference in strain persistence (Fisher test *p*-value > 0.05) (Figure 1a, Table S9). As *B. longum* subsp. *longum* exhibited higher sex-biased stability/persistence in the infant gut, we extended the above-described analysis of dominant strains by exploring the whole *B. longum*

subsp. *longum* strain composition throughout the weaning phase by using different bioinformatic approaches. For this purpose, we obtained 377 non-redundant *B. longum* subsp. *longum* genomes by assembling metagenome-derived data from the infant longitudinal datasets (258 infants, 126 females and 132 males). This genome collection was integrated with publicly available *B. longum* subsp. *longum* chromosomal sequences and, following completeness assessments and dereplication, was employed as a genome database for the inStrain tool[33]. Considering the infants detected by inStrain with cross-weaning persistence of *B. longum* subsp. *longum* at the species level (71 females and 65 males), we observed that at least one strain found in the pre-weaning age was maintained in the post-weaning phase in 52 females (76 %) and 31 males (47 %) (Fisher test, *p*-value = 0.003) (Figure 1b), thus corroborating the findings reported above.

Moreover, after a persistence event (n=83), the early-engrafted *B. longum* strains reached an average relative abundance of 71 % ± 21 % in the overall post-weaning *B. longum* subsp. *longum* strain communities (Figure 1c). These figures, although similar between infant males and females, were significantly higher than those calculated for the *B. longum* subsp. *longum* strains not involved in persistence episodes (n=171, average relative abundance of 29 % ± 29 %, Mann-Whitney test, *p*-value = 0.001) (Figure 1c). These findings imply that when a persistence event occurs, it involves strains ecologically favored to colonize the infant gastrointestinal tract at higher relative abundance, thus dominating the conspecific strain population. Notably, analysis of the overall *B. longum* subsp. *longum* strain population showed that persistence of specific strains seems to occur at a statistically significant higher rate in female infants compared to male infants, also highlighting a superior colonization capability of the persistent *B. longum* subsp. *longum* members compared to the coexisting non-persistent strains.

Based on current scientific data, one may argue that following initial inoculation by the maternal fecal bacteria, diet ingredients such as host-derived glycans play a role in selecting persistent (bifido)bacterial strains, such as *B. bifidum* and *B. longum* subsp. *longum* members, based on their ability to metabolize these amino sugars[34,35]. Such bifidobacterial strains first forage on lactose, HMOs, and perhaps other milk-associated glycans or glycoproteins, subsequently taking advantage of intestinal mucin glycans both as binding sites and as a carbon source in a strain-specific manner, leading to their long-lasting persistence [35].

Identifying genes associated with sex-specific microbial persistence

As *B. bifidum* and *B. longum* subsp. *longum* strains exhibited sex-specific resilience across infant weaning, we performed a comparative genomic analysis to explore the potential genetic features distinguishing these bifidobacterial species from *B. breve* and *B. pseudocatenulatum*, which displayed no or poor persistent behavior or sex-related differences (Figure S1). To gain an accurate overview, the analysis encompassed all complete and well-annotated genome assemblies available from public repositories for a total of 119 genomes belonging to these bifidobacterial taxa. From these surveys, a total of 14 protein families were identified within the core/accessory gene repertoire unique to *B. bifidum* and *B. longum* subsp. *longum,* while they were absent in *B. breve* and *B. pseudocatenulatum* chromosomes (See supplementary text for details) (Figure 1b, Table S10). Among these, we found two GH-encoding genes that were selected for further exploration considering the known importance of host glycans in modulating host-microbe interactions, i.e., the predicted extracellular membrane-anchored mucin-degrading glycosyl hydrolase family 101 (GH101)[34], and the glycosyl hydrolase family 136 (GH136) which is predicted to act as an extracellular lacto-N-biosidase[36] (Figure S5). Specifically, genes encoding GH101 and GH136 enzymes were found in all screened *B. bifidum*

genomes, while 89% and only 38% of the chromosomes belonging to *B. longum* subsp. *longum* showed the presence of genes specifying GH101 and GH136, respectively (Figure 1b; Figure S6; Table S10). The involvement of GH101 and GH136 activities in host glycan metabolism was verified through a transcriptomic survey on *B. bifidum* PRL2010 cultivated under *in vitro* conditions, showing the up-regulation of these GHs when mucin was used as the sole carbon source rather than glucose (MRS-based medium) (Table S11) (See supplementary text for details). These findings suggest that the GH101 and GH136 enzymes, acting on mucin glycan core structures, are involved in the observed long-term colonization in the host gut of *B. bifidum* and *B. longum* subsp. *longum* strains by providing an endogenous source of nutrients in the absence of dietary glycans.

Sex-specific gut persistence of *B. longum* and *B. bifidum* strains from birth to later stages of host life

To validate the hypothesis that *B. bifidum* and *B. longum* subsp. *longum* strains can achieve higher colonization in the human (female) gut across human lifespan, a total of 12,415 cross-sectional metagenomic fecal samples (of which 6,545 or 54 % were of female origin and 5,870 or 46 % were derived from males) from roughly 3,000 infants (0-4 years old, 1,456 female and 1,541 males), 918 children (5-18 years old, 434 females and 484 males), 6,147 adults (19-55 years old, 3,379 females and 2,768 males), and 2,353 elderly (56-90 years old, 1,276 females and 1,077 males) were subjected to strain-level analyses employing the same approach and involving the above described *B. longum* subsp. *longum* and *B. bifidum* reference genome databases in order to analyze the longitudinal infant data (Figure S1). As expected, suckling infants showed the highest prevalence of *B. longum* subsp. *longum* and *B. bifidum* species, exceeding 70 % and 40 %, respectively (Figure 2a, Table S10). Interestingly, when we explored adult populations, we observed that sex markedly

impacted the prevalence of these bifidobacterial taxa (PERMANOVA R^2 = 0.0627 and 0.0558; F = 494.18 and 420.03, both *p*-values = 0.0099; Table S12). Specifically, 41 % and 28 % of the adult female subjects harbored members of *B. longum* subsp. *longum* and *B. bifidum* species, respectively, which, conversely, colonized only 26 % and 11 % of the age-matched male individuals (Fisher test, *p*-values < 0.001) (Figure 2a, Table S10). In contrast, elderly subjects showed the lowest prevalence of bifidobacterial species, ranging from an average of 26 % for *B. longum* subsp. *longum* to an average of 9 % for *B. bifidum* species, with similar values between female and male individuals (Figure 2a, Table S10). Possible confounding factors such as dairy food consumption and lactase persistence were accounted for the differences in *B. longum* subsp. *longum* and *B. bifidum* prevalence between adult females and males. Notably, since this type of information was not available in public datasets, we used geographic regions as proxy variables (south/north Europe for lactase persistence and Europe/China for dairy food consumption), evidencing no significant association between geographic region and sex-dependent prevalence of the target (bifido)bacterial species (Fisher test *p*-values > 0.05) (Supplementary Table S13).

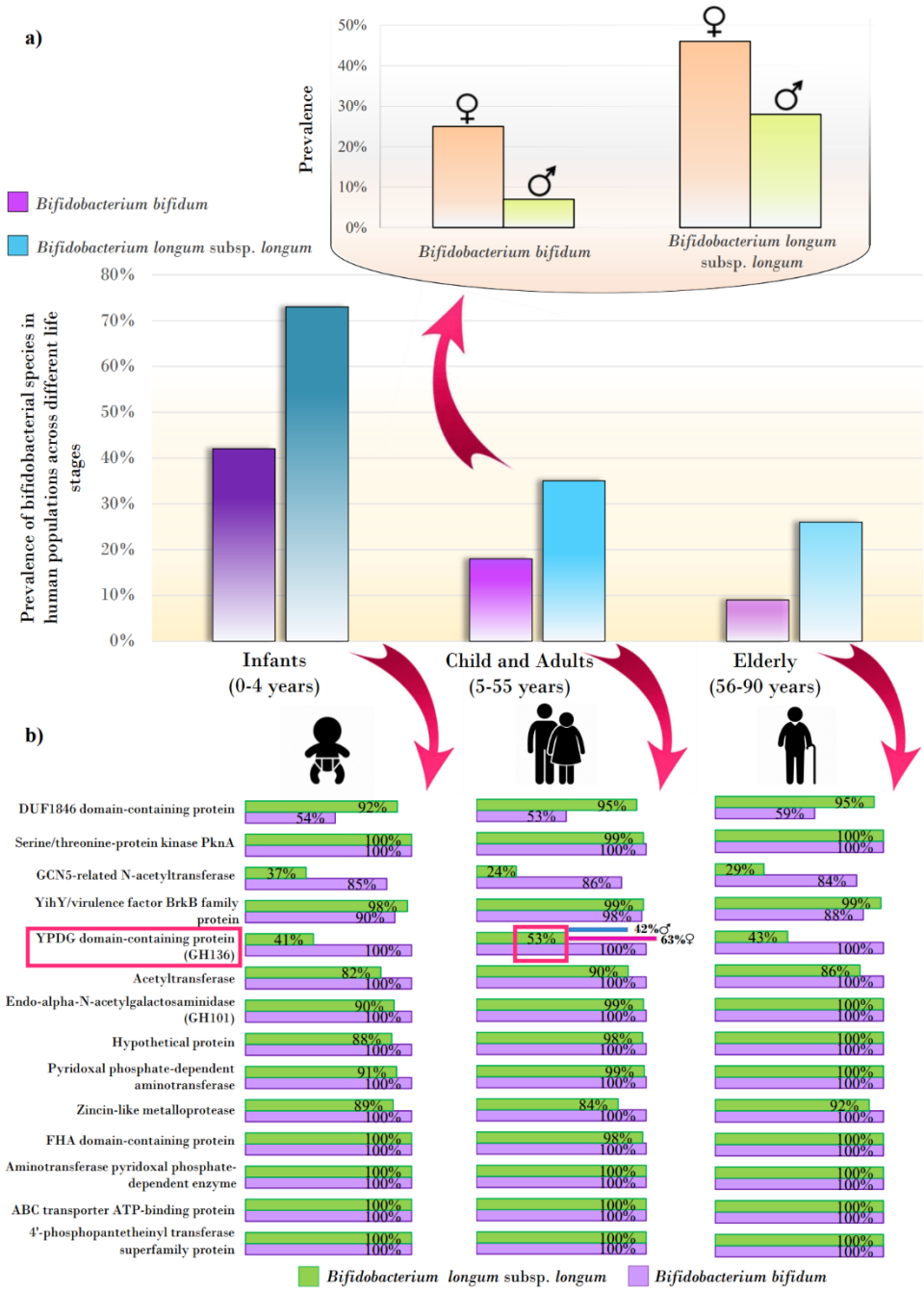


Figure 2. Persistence of *B. bifidum* and *B. longum* subsp. *longum* across human life stages. In panel (a), the vertical bar graph depicts the persistence of *B. bifidum* and *B. longum* subsp. *longum* across the human life span. On the top, bar plots show the differences in the persistence of these bifidobacterial species between female and male individuals aged 5–55 years. In panel **(b)**, horizontal bar charts represent the prevalence of the 14 genes shared by *B. bifidum* and *B. longum* subsp. *longum* across populations. Statistically significant gender-

based differences in the occurrence of GH136 are highlighted with a red frame, and the respective prevalence (percentage) are reported alongside (Fisher test, *p*-value < 0.05).

These findings support the intriguing notion that *B. bifidum* and *B. longum* subsp. *longum* can stably colonize the human gut from infancy to adulthood with an apparent preference in women during their reproductive age, possibly as potential reservoirs for microbial transmission to new generations.

The genomes of *B. bifidum* and *B. longum* subsp. *longum* with completeness equal to or greater than 90 % detected within the infant, child, adult, and elderly gut microbiomes were subjected to a genome-wide screening to assess the occurrence of genes encoding predicted GH101 and GH136 enzymes and thus to deduct their prevalence across populations (Figure 2b, Table S10). According to the survey results, the GH101 of *B. bifidum* and *B. longum* subsp. *longum* were detected between 88 % and 100 % of the 12,415 fecal metagenomic datasets, which, being greater than expected from public genome screening, supports their potential key role in colonization and survival of the human gastrointestinal tract across the entire host lifespan (Table S10).

Instead, the GH136 was found in an average of 45 % of metagenomic samples, with the highest frequency found in adult women (63 %), which is 50 % higher than that observed in age-matched males (42 %) (Fisher test, *p*-value < 0.05) (Table S10). However, such sex-specific differences were not evident in early infancy and older adulthood. Intriguingly, the higher colonization performances observed in females compared with males seem to disappear in individuals older than 50 years, which is concordant with menopause age (Table S10). Recently, sex hormones have been regarded as potent drivers of sexual dimorphism at the gut microbiome level, being associated with compositional differences between sexes and profound changes in the gut microbial community of pregnant women[37–40]. It has been proposed that,

besides its role in the modulation of the immune system and bile acid secretion, which may then regulate the gut microbiota, steroid sex hormones can be metabolized by specific gut-associated bacterial enzymes, thus directly impacting microbial metabolism and growth[37,41,42]. Although little is yet known about the impact of sex on the human gut microbiota and even less about the underlying mechanisms, it has been reported that a higher level of sialylation characterizes the intestinal mucus of the female gut compared to that of the male population[43]. Furthermore, it has also been argued that the female sex hormone estradiol may up-regulate the expression and glycosylation of human mucins[44–46], possibly shaping the female gut microbiota in favor of mucin-utilizing (bifido)bacteria. Interestingly, the estradiol level is higher in females than males, even in the prepubertal phase[47]. Indeed, the hypothalamic-pituitary-gonadal axis, which is involved in the development and regulation of the reproductive system, undergoes transient activation during the first six months of life in males and the first two years in females[48,49]. This event, called minipuberty, or “endocrine puberty", induces testosterone production in males and estradiol in females[50–52].

In addition to these hormone-driven effects, it should be noted that complex mechanisms of DNA methylation patterns in colon tissue-specific genes (excluding loci located in the X and Y chromosomes) may contribute to the establishment of sex- and age-related differences in the intestinal environments[53,54]. Consistently, variability in methylation signatures was observed in subjects of various ages, including both newborns and adults, when comparing females to males[53,55].

Altogether, these findings suggest that, compared to males, the female-specific intestinal surroundings, including glycan structures, may be structurally more predisposed to creating a suitable environment for certain early colonizing microbial species, such as *B. bifidum* and *B. longum*. In particular, when the gene encoding the GH136 enzyme is present, it could play a part in enhancing (bifido)bacterial

persistence in women of reproductive age, possibly due to a sex-specific mucus structure.

The role of GH136 in bacterial vertical transmission from mother to newborn.

B. bifidum and *B. longum* represent two of the main species that can be maternally inherited by means of vertical transmission[56]. In particular, colonization of the infant gut from maternal gut-associated strictly anaerobic species, such as *Bifidobacterium*, is believed to occur through direct contact with maternal gut microbiota during birth and/or involve an entero-mammary pathway, which may transfer ecologically well-adapted bacteria via breastmilk to infants[57–60]. Starting from this scientific evidence and exploiting the non-ubiquitous presence of GH136-encoding genes in *B. longum* strains, we decided to estimate the involvement of GH136 in mother-to-infant vertical transmission events through inspection of publicly available metagenomic fecal samples from 132 full-term vaginally delivered healthy newborns and their corresponding mothers[61,62] (Figure S1). Notably, fecal samples collected from mothers at childbirth and newborns at one month were subjected to *B. bifidum* and *B. longum* subsp. *longum* strain tracking analyses, and the reads were then mapped to the GH136 gene sequence.

Our results indicated that specific strains of *B. longum* subsp. *longum* and *B. bifidum* were present in the samples from mother-infant dyads in 36.4 % and 29.5 % of the screened cases, respectively, suggesting perinatal vertical transmission events. Interestingly, 72.9 % of the *B. longum* subsp. *longum* strains detected as vertically transmitted from mothers to newborns harbour the GH136-encoding gene, while just 28.6 % of the *B. longum* subsp. *longum* strains that did not appear to be involved in vertical transmission events were shown to possess the GH136 gene (Fischer Test p-value < 0.001, Table S14). Accordingly, sex-related genetic and epigenetic host factors seem to be associated with specific microbial genomic features to ensure

persistence and, therefore, establishment of selected consortia of bifidobacterial strains that may be maternally transmitted through delivery to the offspring.

Retrospective clinical studies support the GH136-driven, long-term gut persistence of *B. longum* subsp. *longum* in women. In order to assess the role of the accessory GH136 in the colonization and persistence of *B. longum* subsp. *longum* in the human gut, we analyzed data from published retrospective clinical studies in which healthy human participants received daily oral doses of viable cells of two genetically different *B. longum* subsp. *longum* strains[63,64] (Figure S1). Specifically, a total of 21 healthy individuals (52% females) received a treatment consisting of a daily dose of 10^{10} viable cells of *B. longum* subsp. *longum* AH1206, a strain that harbours genes encoding the GH136 and GH101 enzymes (Table S10). A second group of 10 individuals (40 % females) consumed the same daily dosage (10^{10} viable cells) of a *B. longum* subsp. *longum* strain named AG1, possessing GH101 but lacking the GH136 gene (Table S15).

Shotgun metagenomic data of stool samples collected before the intervention (baseline), after 21-28 days of oral bacterial administration (treatment), and after the follow-up period (persistence) were exploited to evaluate the persistence of *B. longum* subsp. *longum* AH1206 and AG1 strains (Figure 3a).

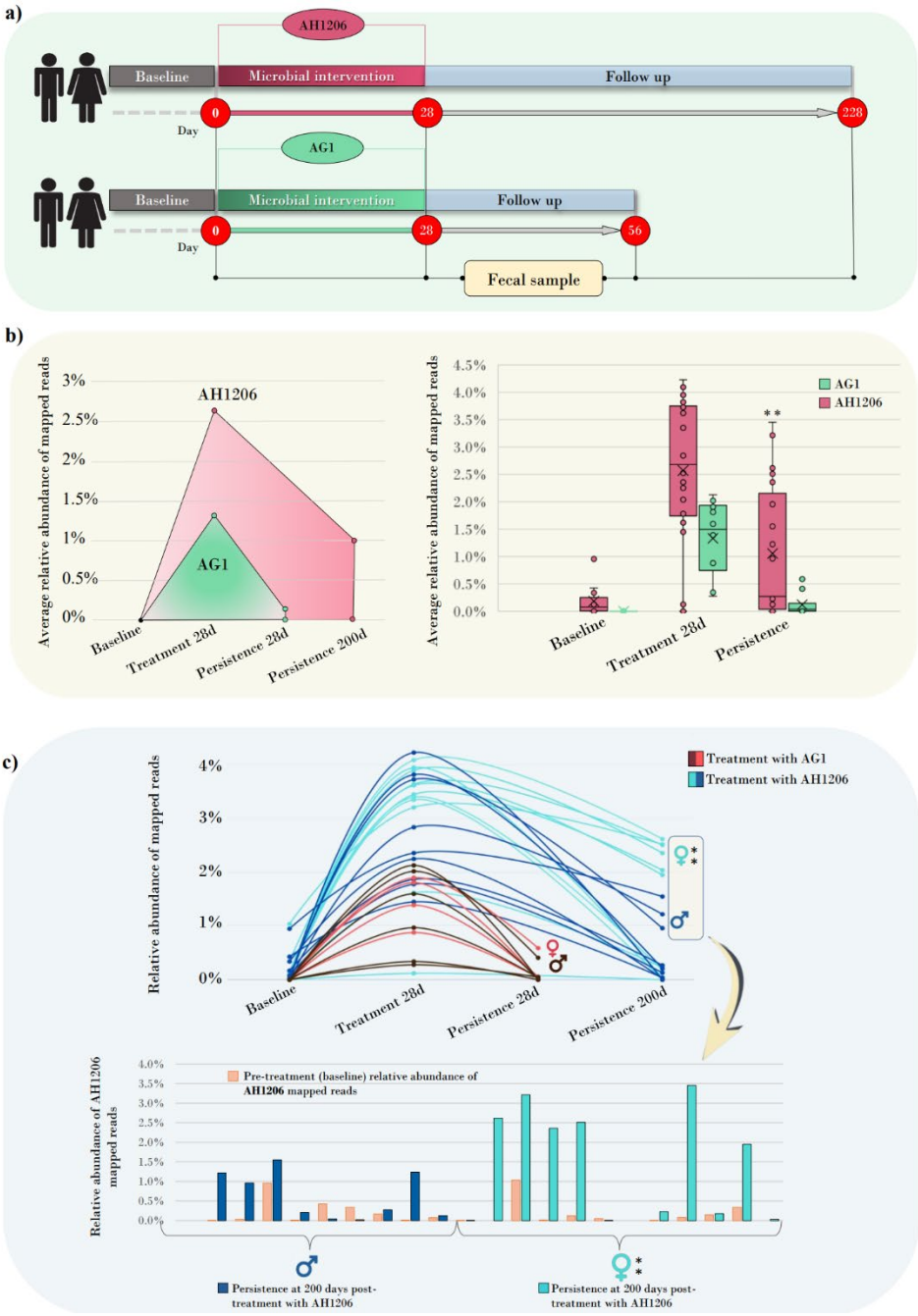


Figure 3. Analysis of data from human retrospective clinical studies based on the supplementation of *B. longum* subsp. longum AH1206 and AG1 strains. Panel **(a)** shows the experimental outline of the human retrospective trials considered in this study. Panel **(b)** reports the average relative abundance of mapped reads of *B. longum* subsp. *longum* AG1206 and AG1 in the metagenomic samples during the bacterial supplementation and follow-up. In panel **(c)**, the significant difference in the average abundance of AH1206 and AG1 mapped reads between follow-up (persistence) and the corresponding baseline is indicated by an

asterisk in the Box and Whisker plot (**p-value < 0.01; Wilcoxon signed-rank test, p-value = 0.009). The boxes are determined by the 25th and 75th percentiles. The whiskers are determined by 1.5 interquartile range (IQR). The line in the boxes represents the median, while the cross marker (X) represents the average. In panel **(d)**, the trend of AG1206 and AG1 mapped reads (relative abundance) across the corresponding interventional studies (n = 22 and n = 10, respectively) is shown for each sample. Differences in the persistence of AH1206 strains between female and male populations (n = 21, 11 females and 10 males) are detailed in the bar plots depicting the average relative abundance of AH1206 mapped reads detected at baseline (n = 21, 11 females and 10 males) and after the termination of treatment (n = 21, 11 females and 10 males) (**p-value < 0.01; Wilcoxon signed-rank test, p-value = 0.039).

Fecal metagenomic reads from each subject were mapped against the administered bifidobacterial strain genome sequences (AH1206 or AG1), considering only > 99% homology matches. The results showed that both AH1206 and AG1 strains were detectable at the end of the 28-day treatment period and then decreased after the termination of microbial supplementation in a strain-specific manner (Figure 3b, Table S15). Remarkably, the relative abundance of AH1206 mapped reads remained significantly higher compared with the pre-treatment baseline even 200 days after completion of treatment (Wilcoxon test, p-value < 0.01) (Figure 3b, Table S15). In contrast, the relative average abundance of AG1 mapped reads was not significantly higher when compared to the situation before the microbial intervention as early as 28 days after interruption of consumption (Wilcoxon test, p-value > 0.05), indicating that the level of AG1 one month after treatment was reverted to that observed in the pre-treated microbiomes. Intriguingly, when assessing the degree of long-term AH1206 colonization among female and male volunteers, we identified gender-related differences in the level of strain persistence. Indeed, at the persistence test time-point (around 200 days of follow-up), the average relative abundance of AH1206-related mapped reads from females was found to be significantly higher compared with their own baseline (Wilcoxon test, p-value < 0.01), while male participants did not exhibit such long-term persistence of AH1206 (paired t-test of

average mapped reads at 200 days vs. baseline, *p*-value > 0.05) (Figure 3c, Table S15).

Overall, these findings possibly indicate that GH136 positively impact the (female host-associated) persistence of *B. longum* subsp. *longum* strains.

Evaluation of the molecular interaction of persistent and non-persistent *B. longum* strains with human intestinal cells through transcriptomics analyses

To assess the role of the non-ubiquitously present *B. longum*-encoded GH136 in host-microbe cross-talk and to investigate molecular interactions between persistent and non-persistent *B. longum* subsp. *longum* strains and the host cells, we applied an *in vitro* approach involving human cell lines placed in contact with bacterial cells (Figure S1). Specifically, we cultivated Caco2/HT29-MTX cell monolayers in direct physical contact with *B. longum* subsp. *longum* PRL2022 or 1898B strains, which had been selected based on the presence or absence of the gene encoding the accessory GH136, respectively. Subsequently, the transcriptomes from both bacterial and human cell lines were investigated through RNA-Seq experiments aimed at evaluating the differentially expressed genes (DEGs) between each treatment (PRL2022- and 1898B-Caco2/HT29-MTX contact) and the respective control conditions (absence of contact), considering statistically significant a fold-change ≥ 2 at a *p*-value ≤ 0.05 after correction for multiple comparisons using the False Discovery Rate (FDR) procedure (Figure S7, Figure S8, Table S16, Table S17).

Following host cell contact, a total of 334 and 492 bacterial genes from PRL2022 and 1898B strains, respectively, were classified as DEGs when compared to control samples (Table S16). Among the statistically significant up-regulated transcriptomes (~50 % of total DEGs) (Figure 4a, Figure S6), we found transcripts corresponding to genes encoding priming and processing glycosyltransferases involved in exopolysaccharide (EPS) production, several carbohydrate and amino acid

modifying enzymes, and various protein kinases (Table S16). Focusing on (bifido)bacterial host-glycan and carbohydrate metabolism genes, it was found that both PRL2022- and 1898B-related transcriptomes showed significantly higher expression of the GH101-corresponding gene (JL750_RS06690, BLSL_RS08555) when compared to control samples, in addition to genes predicted to encompass a number of 16-19 ABC transporters involved in carbohydrate uptake, including that identified above among the 14 genes shared between *B. longum* and *B. bifidum* (Figure 4a, b, d, e; Figure S7). Moreover, the GH136 gene (JL750_RS09800), present only in *B. longum* subsp. *longum* PRL2022, was shown to be up-regulated (Figure 4b, Table S16).

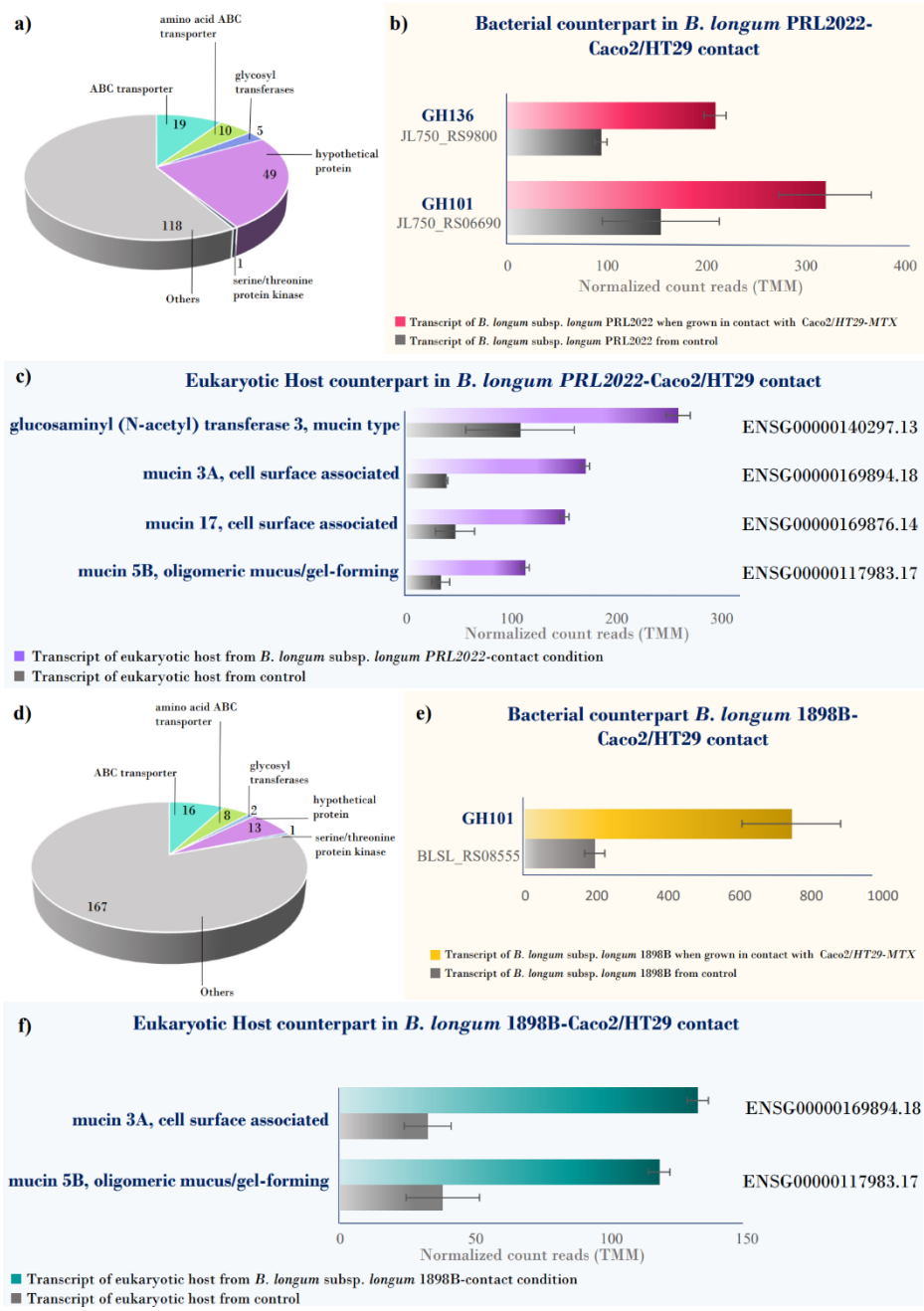


Figure 4. Focus on differentially expressed bacterial mucin-degrading genes and host mucin-producing. In panel **(a)**, the chart graph highlights the number of different gene functional categories among the up-regulated genes of *B. longum* subsp. *longum* PRL2022 when grown in contact with Caco2/HT29-MTX host cells. In panel **(b)**, the horizontal bar plot shows the differentially expressed GH101 and GH136 mucin-degrading genes between *B. longum* subsp. *longum* PRL2022 grown in contact with Caco2/HT29-MTX human cells and control. In panel **(c)**, the horizontal bar plot reports the mucin-producing genes found up-regulated in the eukaryotic host transcriptome from *B. longum* PRL2022-exposed

Caco2/HT29-MTX vs. control. Panel **(d)** depicts the number of different gene functional categories among the upregulated genes of *B. longum* subsp. *longum* 1898B when grown in contact with Caco2/HT29-MTX host cells. Panel **(e)** shows the differentially expressed GH101 mucin-degrading genes between *B. longum* subsp. *longum* 1898B grown in contact with Caco2/HT29-MTX human cells and control. In panel **(f)**, the horizontal bar plot reports the mucin-producing genes found up-regulated in the eukaryotic host transcriptome from *B. longum* 1898B-exposed Caco2/HT29-MTX vs. control. Error bars represent standard deviations from three independent replicates. After normalization of row counts, genewise exact tests were computed to assess the differential expression of each gene. Adjustment of *p*-values for multiple hypotheses was performed through the false discovery rate (FDR) procedure.

As a result, intestinal bacterial colonization appears to be a multifactorial process that involves various carbohydrate modifying enzymes and microbial surface components. These appear to corroborate the involvement of the (bifido)bacterial mucin-degrading enzymes GH101 and GH136 in host-microbe interplay and mucosal surface colonization. However, as mammalian hormones may affect bacterial gene expression[65–67], we evaluated whether the endocrine milieu contained in the fetal bovine serum (FBS, used in combination with DMEM for human cell culture, see Materials and Methods section) could impact the expression of the GH136 gene. Specifically, we differentially grew PRL2022 on the culture medium DMEM with and without FBS, obtaining no significant changes in RT-qPCR-based GH136 gene expression results (t-test p-value > 0.05).

Furthermore, the transcriptome of human cell lines placed in contact with *B. longum* subsp. *longum* was assessed and compared with that achieved in the absence of bacterial cells. Notably, a total of 1253 (874 up-regulated) and 1404 (892 up-regulated) host cell-related genes were identified as DEGs in the PRL2022-exposed and 1898B-exposed groups, respectively (Figure S8, Table S17). Among the *B. longum*-induced up-regulated host transcripts, we found genes encoding pattern recognition receptors, which react to bacteria (e.g., Toll-like receptors)[68,69], and cell

signaling molecules that aid cell communication in immune responses (e.g., cytokines)[70] (Table S17).

Focusing on the expression of human mucin genes, transcriptome analysis of Caco2/HT29-MTX cells upon exposure to *B. longum* subsp. *longum* PRL2022 (which harbours a GH136-encoding gene) revealed up-regulation of genes encoding mucin5B (mucus/gel-forming), mucin3A, and mucin17 (cell surface-associated)[71], as well as glucosaminyl (N-acetyl) transferase 3, which catalyzes the formation of core 2 and core 4 O-glycans on mucin-type glycoproteins[72,73] (Figure 4c, Table S17). Remarkably, only mucin3A and mucin5B were expressed at significantly higher levels in the non-persistent *B. longum* subsp. *longum* 1898B-exposed host-derived transcriptome compared with control samples (Figure 4f, Table S17).

Overall, these findings suggest that *B. longum* subsp. *longum* strains significantly influence the transcriptome of human intestinal cells, modulating the expression of genes involved in the synthesis of mucus layer components. Remarkably, *B. longum* subsp. *longum* PRL2022 (encoding the GH136 enzyme) appeared able to stimulate the expression of host mucin and mucin-related genes to a greater extent (100 % increase in the number of up-regulated host mucin-related genes) than the conspecific 1898B strain lacking the GH136 gene, implying a possible competitive advantage and enhanced persistence in the gut of (female) human hosts.

Conclusions

Several studies have demonstrated that vertical transmission from mother to infants is a pivotal route for early establishment of (certain) members of the gut microbiota, which can persist across subsequent stages of human life, although with a reduced abundance[8,28]. In this context, it is particularly important for the human female host to sustain the persistence of those early colonizers that may later be maternally transmitted to new generations.

In the current study, through multi-omics approaches, we revealed that *B. bifidum* and specific strains of *B. longum* subsp. *longum* can establish a long-lasting colonization behavior preferentially in the intestinal environment of females compared with male individuals. Notably, these taxa are renowned for their highest level of vertical transmission among bifidobacterial species[22,74]. Interestingly, screening for the genetic determinants possibly involved in (female) human gut persistence led to the identification of two mucin-degrading GH families, GH101 and GH136, present in all *B. bifidum* and particular *B. longum* subsp. *longum* strains.

Specifically, the non-ubiquitous distribution of the GH136 gene within *B. longum* subsp. *longum* taxon highlights the involvement of this gene in mother-to-infant vertical transmission and led to its identification as a strain-specific genetic key determinant for stable female gut colonization. Recent studies have considered the importance of sex hormones in sex-dependent trajectories of the gut microbiome, besides the well-known environmental factors, age, dietary habits, geographical origin, and antibiotics[37–40,42]. Consistently, a higher level of sialylation, likely driven by sex-related hormones and potentially complex epigenetic mechanisms, was previously noted in the intestinal mucus of the female gut compared with that of the male population[43–46], suggesting that a sex-specific intestinal environment could be in a prime location to actively select persistent mucin-degrading (bifido)bacterial colonizers.

Overall, our findings propose an intriguing and novel strict cooperation between host physiology and microbial genetics as a result of ancient (bifido)bacteria-human coevolution aimed at ensuring the maternal persistence of those microbial species that may at some point be vertically transferred to the next generation.

MATERIALS AND METHODS

Study population

The study included 11 vaginally delivered infants born after an uncomplicated pregnancy and recruited at the Central University Hospital of Asturias (Northern Spain). The study was approved by the Regional Ethical Committee of Asturias Public Health Service (Ref.Nº 51/18) and the Ethical Committee of CSIC (Ref 136/2018). Informed written consent was obtained from each infant's parent. Fecal samples were collected at scheduled appointments from one month to two years after birth (Table S1). After sequencing (see below), the number of persistent strains detected in each post-weaning sample (corresponding to 12 or 24 months after birth) was normalized to account for differences in sequencing depth. Specifically, a normalizing factor was calculated for each sample as the ratio between the mean sequenced reads across the whole dataset and the number of sequenced reads in the sample. The obtained normalizing factor was then applied to the number of persistent strains identified in the metagenomic sample, thus obtaining an adjusted number of persistence strains.

To corroborate the findings observed in the study cohort, a validation cohort was constructed employing a large and deep-sequenced publicly available longitudinal infant dataset[29] (PRJEB32631) (Table S1). Specifically, infant fecal samples collected in the first 21 days after birth (pre-weaning phase) and after six months postnatally were selected as suitable for validating the (bifido)bacterial persistence pattern (Table S1). This validation dataset was parsed by combining the recent mGEMS and mSWEEP methods[30,75–77], which assign short metagenomic reads to genomic bins corresponding to individual genomes of predefined species and estimate the relative abundances of reference bacterial strains, respectively. The k-mer-based pseudo-alignments against pre-build reference sequence databases were

obtained through the Themisto software (version 3.1.2). Notably, this procedure was also employed to recover a set of reference genomes belonging to the target species characterizing the pre-weaning infant gut microbiome. In detail, a total of 371 MAGs belonging to *B. bifidum*, *B. longum* subsp. *longum*, *B. breve*, *B. pseudocatenulatum*, and *E. coli* were successfully reconstructed with an average completeness degree of 86.7 % ± 10 % (based on checkM method) and were employed as reference genomes to confirm the within-host strain variation patterns observed throughout the weaning phase in the infant study population.

Publicly available datasets

To extensively investigate the host sex-related bacterial persistence pattern, the newly sequenced longitudinal dataset was supplemented by 357 publicly available shotgun metagenomic samples from two studies that sampled 113 infants longitudinally[11,32]. In particular, we selected datasets corresponding to the analysis of the healthy infant gut microbiome at pre- (< 6 months old) and post-weaning (> 6-12 months old) development stages. Additionally, to create a comprehensive population-based cohort covering different age groups, cross-sectional fecal metagenomic samples of 12,415 healthy individuals aged from a few days after birth to 90 years were retrieved from 146 different publicly accessible studies. Specifically, only healthy (control) individuals and a single sample per each individual were retained from every study. A complete list of samples and metadata is available in the supplementary material (Table S1). Notably, while this collected population dataset does not include subjects for whom any parental relationships emerged from corresponding metadata, shotgun data from mother-infant dyads were considered separately. Batch effects caused by different sequencing technologies were controlled by selecting only raw data produced through Illumina DNA sequencing platforms. Moreover, to overcome the potential confounding effects of library size, we randomly selected 5,000,000 reads

from each sample, such that all samples have the same library size, while samples with fewer reads than 5,000,000 were discarded. Before applying this procedure, we used 250 deep-sequenced metagenomic samples (PRJEB32631) to compare the results obtainable by strain-level profiling using all the available reads (average of 8,826,752 ± 1,931,731 after *Homo sapiens* filtering) and a random subsample of 5,000,000 reads to ensure that no valid information on the microbial community structure was lost. The Mann–Whitney U test calculated on the strain profiles of four target species revealed comparable ability in detecting within-species variation in the metagenomic reads (Table S8).

Confounding variables related to population structure, such as ethnicity/geographical origin and age, were tested through the permutation analysis of variance (PERMANOVA). Specifically, we stratified by age groups and assessed the statistical significance (p-value), the proportion of explained variance (R^2), and the effect size (F value) for each categorical variable. In addition, confounding effects of possible dairy food consumption and lactase persistence were tested by assuming geographical region as a proxy parameter. Specifically, subsets of metagenomic samples from Southern Europe (n=417, 217 males and 200 females) and Northern Europe (n=413, 186 males and 227 females) were selected to test the association between sex-dependent prevalence of *B. longum* subsp. *longum* and *B. bifidum* and lactase persistence, while fecal samples of subjects from China (n=831, 413 males and 418 females) were compared with those from Europe (n=830, 403 males and 427 females) to account for dairy consumption.

Bacterial DNA extraction

Stool samples were stored on ice immediately after collection and shipped to the laboratory under frozen conditions, where they were preserved at –20°C until processing. DNA extraction was performed using the QIAmp DNA Stool mini-kit

according to the manufacturer's instructions (Qiagen, Germany). DNA quantification was achieved employing the Qubit fluorometer (Thermo Fisher Scientific).

Whole-genome sequencing and taxonomic classification

Shotgun metagenomic sequencing was performed by GenProbio srl (www.genprobio.com). DNA library preparation was performed using the Nextera XT DNA sample preparation kit (Illumina, San Diego, CA) according to the manufacturer's instructions. One ng input DNA from each sample was used for library preparation. The isolated DNA underwent fragmentation, adapter ligation, and purification. The ready-to-go libraries were pooled equimolarly, denatured, and diluted to a sequencing concentration of 2 pM. Sequencing was performed on a NextSeq 500 instrument (Illumina, San Diego, CA), according to the manufacturer's instructions, using the 2x150 bp High Output sequencing kit and spike-in of 1 % PhiX control library. Whole-metagenome shotgun (WMGS) sequencing of the abovementioned 43 infant gut microbiomes produced an average of 7,940,484 ± 2,078,529.565 paired-end 150 bp reads per sample. Following quality filtering (minimum mean quality score, 20; window size, 5 bp; and minimum length, 80 bp) and removal of reads that map on the *Homo sapiens* genome, an average of 6,355,063 ± 1,425,089 microbial reads per sample were retained (Table S1).

Taxonomic profiling of sequenced reads, including those retrieved from publicly available shotgun datasets, was achieved with the METAnnotatorX2 bioinformatics platform[78], using the up-to-date RefSeq (genome) sequence database retrieved from the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/refseq/). Species-level taxonomic classification of each read was achieved through Megablast[79] (with option -e-value 1e-5, -qcov_hsp_perc 50) using > 94% alignment identity. Reads that showed the same sequence identity against more than one bacterial species were discarded. Similarities

between samples (beta-diversity) were computed by Bray-Curtis dissimilarity based on species abundance. PCoA representation of beta-diversity was performed using ORIGIN 2021 (https://www.originlab.com/2021).

Bacterial genome assembly

For each unique host infant, representative genomes of 11 main gut-associated microbial species (Table S3) were reconstructed as previously reported[17,78], leading to *de novo* metagenomics assembly and taxonomic classification of 67 bacterial genomes with a minimum average read coverage of 12X and a total genome size compatible with what was reported in literature. In detail, raw data of shotgun metagenomic sequencing (fastq files) that passed quality filtering and human genome mapping were used as input for SPAdes assembler v3.12[80], using default parameters enabling the metagenomic flag option (-meta). SPAdes parameters were combined with minimum k-mer sizes of 21, 33, and 55 to a maximum of 77, 99, and 127 based on the paired-end read length, as previously described[81]. Following assembly, ORFs of each assembled contig were predicted with Prodigal[82] with default parameters and then annotated using MEGAnnotator software[83,84].

Construction of reference microbial genome databases and metagenomic strain-level analyses

To untangle strain communities and investigate strain dynamics in the infant, adult, and elderly gut microbiomes, 11 species-specific databases of reference genomes were constructed using 63 MAGs coupled with the genome sequences publicly available on the NCBI RefSeq database (complete and draft high-quality genome sequences with less than 90 contigs) (Table S4). Specifically, for each species, collected genomes (MAGs and publicly available) were processed through the open-source software Strain Genome Explorer (StrainGE)

(github.com/broadinstitute/strainge) and dereplicated (clustered) according to the Average Nucleotide Identity (ANI) values (threshold of 99 % ANI value) (Table S4). Subsequently fastq from shotgun metagenomic sequencing were analyzed using the Strain Genome Search Tool (StrainGST, an integrated component of the StrainGE tool suite)[85] that, following k-merizing both input fastq and reference genomes (straingst kmerize -k 23), iteratively compares the sample-associated k-mers set with those obtained from the reference genomes. As a result, StrainGE returns the reference genomes more similar (strongest StrainGST score, default threshold of 0.02, which is optimized to maximize sensitivity and minimize false positives) to those present within the metagenomic samples. This approach was used to investigate longitudinal strain stability in infants by comparing the strains profiled at 0-6 months (pre-weaning) with those identified at 12-24 months (post-weaning). To compare the whole *B. longum* subsp. *longum* strain communities between infant females and males throughout the weaning phase, we metagenomically assembled an additional 542 *B. longum* subsp. *longum* genomes from the 113 publicly available infant metagenomes (357 longitudinal samples) as described above. This collection was enriched with 404 *B. longum* subsp. *longum* genomes (complete and draft genomes with less than 90 contigs) retrieved from public repositories. Subsequently, the obtained 946 *B. longum* subsp. *longum* chromosomes were processed through the DRep[86] and checkM[87] tools to cluster essentially identical genomes (ANI values = 99 %) and select high-quality reference genomes from each replicate set. As a result, we obtained a non-redundant database of 277 *B. longum* subsp. *longum* strains, which was used for processing metagenomic fastq with inStrain tool under default parameters [33]. We used the value obtained by normalizing the genome coverage on the corresponding genome length as a proxy of the relative abundance of each *B. longum* subsp. *longum* reference strains.

qRT-PCR analyses

To validate the highlighted strain dynamics of the infant gut microbiome, we performed a qRT-PCR-based assays focused on 3 cases in which a marked switch of the dominant genetic variant over time was observed.

Firstly, strain-specific primers were designed to target unique genetic sequences of the predicted persistent strains (Table S6). The strain specificity of each designed primer pair was assessed *in silico* through primer-BLAST and *in vitro* validated by end-point PCR reactions performed by using the following thermal cycling protocol: 5 min at 94 °C for one cycle, followed by 30 cycles at 94 °C for 30 s, primer pair-specific annealing temperature for 30 s and 72 °C for 50 s, and a final cycle of 72 °C for 5 min. For this validation step, each PCR reaction was performed on the DNA extracted from each sample of the longitudinally collected fecal samples from a specific infant, together with DNA extracted from other taxa belonging to the same species of the targeted strain. Once the primer strain-specificity had been confirmed, a second end-point PCR was carried out using the DNA extracted from the fecal sample showing the higher average relative abundance of the strain containing the unique genetic sequence target of the designed primers. The obtained amplicon was then purified using the NucleoSpin PCR & Gel Clean Up kit (Macherey-Nagel, France), following the manufacturer's guidelines. The purified amplicon was subsequently used in a qRT-PCR run as the standard DNA to build a standard curve since no chromosomal DNA was available. Specifically, each qRT-PCR reaction mix contained 7.5 µl 2x SensiFast Sybr No-Rox kit (Meridian Bioscience, USA), 5 µl of DNA diluted to 10 ng/ µl, each of the forward and reverse primer at 0.5 µM and nuclease-free water was added to obtain a final volume of 15 µl. Each qRT-PCR run was carried out on a CFX96 system (BioRad, CA, USA) using the following protocol: 95 °C for 2 min, followed by 40 cycles of 95 °C for 5 s and 60 °C for 30 s, and a melting curve from 65 °C to 95°C with increments of 0.5 °C/s. Negative

controls (no DNA) for each primer set were included in each run, while standard curves were built using the CFX96 software (Biorad).

Glycobiome prediction

The metagenomically-assembled bifidobacterial genomes predicted to be either persistent or transient in the infant gut were screened for genes encoding the catalytic hydrolysis of the glycosidic bond. Prediction of glycosyl hydrolase (GH)-encoding genes and their classification into GH families were achieved through similarity search in the carbohydrate-active enzyme (CAZy) database[88] (BLAST cutoff e-value of 1×10^{-10}).

Comparative genomic analyses

Only complete genome sequences belonging to *B. bifidum*, *B. longum, B. breve,* and *B. pseudocatenulatum* were retrieved from NCBI's RefSeq genome database and subjected to core-genome analysis using the Pangenome Analysis Pipeline (PGAP) v1.1 (--identity 0.5 --coverage 0.8 --cluster --method GF)[89]. Specifically, the predicted proteome of each bifidobacterial genome was screened for orthologues against the proteome of the other conspecific strains by BLAST analysis (cutoff e-value $< 1 \times 10^{-5}$ and 60% identity over at least 80% of both protein sequences). The resulting data were cataloged into functional gene clusters, also designated as Cluster of Orthologous Groups (COGs), employing MCL (graph-theory-based Markov clustering algorithm)[90], through the gene family (GF) method (cutoff e-value of $1x10^{-10}$). A pan-genome profile was built using the algorithm provided as part of the PGAP software, based on a presence/absence matrix encompassing all COGs identified in the considered genomes. Accordingly, the protein families shared between *B. bifidum* and *B. longum* and absent in the *B. breve* and *B. pseudocatenulatum* genomes were collected.

Genetic characterization of the shared *B. bifidum* and *B. longum* genetic traits and evaluation of their occurrence in infant, adult, and elderly populations

Protein sequences of the identified 14 COGs shared by *B. longum* and *B. bifidum* were subjected to extensive homology searches, and domain and localization predictions. In detail, Pfam v34.0 (https://pfam.xfam.org/), InterPro 86.0 (https://www.ebi.ac.uk/interpro/), and the Simple Modular Architecture Research Tool (SMART) (http://smart.embl-heidelberg.de/)[91] were employed to identify the protein domains, while SignalP v5, Psort v3.0.3, and THMM v2 were used for cellular localization prediction. SWISS-MODEL (https://swissmodel.expasy.org/) online tool was for 3D protein structure models and comparative modeling[92]. BLASTP and tBLASTN search with an E-value of $1e^{-5}$ was performed against the integrated non-redundant protein sequence data resources (nr) for functional annotation of the coding sequences. BLASTP analysis (E value cutoff of $1e^{-5}$) was also used to screen for PDC homologs in the *B. bifidum* and *B. longum* genomes identified in the multi-population metagenomes. Moreover, each of the 14 identified PDCs was aligned with the WGS reads to determine their prevalence and abundance in the population cohort, as previously described[93]. Briefly, following quality filtering (minimum mean quality score, 20; window size, 5 bp; quality threshold, 25; and minimum length, 100 bp) and removal of reads that map on the *Homo sapiens* genome, the final mapping against the 14 PDCs was performed using Bowtie2[94] through multiple-hit mapping and "very-sensitive" policy. The mapping was performed using a minimum score threshold function (–score-min C,−13,0) to limit reads of arbitrary length to two mismatches and retain those matches with at least 99% full-length identity. HTSeq software[95] (running in union mode) was employed to calculate read counts corresponding to each PDC gene.

Analysis of fecal metagenomic samples of 132 mother-infant pairs

A total of 164 metagenomic fecal samples from mothers and their healthy term vaginally derived newborns were retrieved from public repositories (PRJEB6456, PRJNA475246). Specifically, we selected DNA sequencing data generated from shotgun metagenomic sequencing using Illumina platforms. For identification of vertical transmission events involving *B. longum* subsp. *longum* strains, we used StrainGE tool on fecal samples from the mother at delivery and newborn at 14-30 days after birth as described above. Additionally, Bowtie2 was employed to map metagenomic reads against the sequence of GH136, and HTseq software was used to compute reads row counts as described above.

Growth of *B. bifidum* PRL2010 on mucin and RNA-Seq analyses

B. bifidum PRL2010 cells were grown at 37°C under anaerobic conditions (2.99% H2, 17.01% CO2, and 80% N2) (Concept 400; Ruskin) in De Man-Rogosa-Sharpe (MRS) broth (Sharlau Chemie, Barcelona, Spain) supplemented with 0.05% (wt/vol) l-cysteine hydrochloride. Viable cells were inoculated in 30 ml of freshly prepared modified MRS without glucose (mMRS) supplemented with 0.5% mucin. Cells were inoculated with an OD_{600nm} of 0.1. After inoculum, growth was monitored, and at an OD_{600nm} between 0.6 and 0.8 (exponential phase), cells were harvested by centrifugation at 6000 rpm for 5 min. Growth assays were carried out in triplicate. Total RNA of *B. bifidum* PRL2010 cultures was isolated as previously described[27]. The quality of the RNA was verified by employing a Tape station 2200 (Agilent Technologies, USA). RNA concentration and purity were evaluated using a spectrophotometer (Eppendorf, Germany). For RNA sequencing (RNA-Seq), from 100 ng to 1 µg of extracted RNA was treated to remove rRNA using the QIAseq FastSelect – 5S/16S/23S following the manufacturer's instructions (Qiagen, Germany). The yield of rRNA depletion was checked using a Tape station 2200

(Agilent Technologies, USA). Subsequently, a whole transcriptome library was constructed using the TruSeq Standard mRNA Sample preparation kit (Illumina, San Diego, USA). Samples were loaded into a NextSeq high-output v2.5 kit (150 cycles, single end) (Illumina) following the technical support guide. Demultiplexed reads were quality filtered (with overall quality and length filters) and aligned to the *B. bifidum* PRL2010 reference genome through BWA[96]. Counts of reads that overlap ORFs were performed using HTSeq software[95]. Analysis of the RPKM values and false discovery rate correction (cut-off 0.01) was performed using DESeq2[97] and the formula RPKM = numReads/(geneLength/1,000 × totalNumReads/1,000,000)[98]. The experiment was conducted in triplicate.

Human cell line trials

Caco-2 cells, derived from a colorectal adenocarcinoma of a human male donor (purchased from ATCC) and HT29-MTX, a human colon carcinoma-derived, mucin-secreting goblet cell line from a female donor (kindly provided by prof. Antonietta Baldi, University of Milan) were cultured in Minimum Essential Medium (MEM) and Dulbecco's Modified Eagle's medium (DMEM) with high glucose (4.5 g/L) and 10 mM of sodium pyruvate, respectively, as previously described[99]. Both media were supplemented with 10% Fetal Bovine Serum (FBS), 2 mM glutamine, 100 μg/ml streptomycin, and 100 U/ml penicillin. Cultures were maintained at 37°C in a 5% CO_2 humidified atmosphere in 10-cm dishes and passaged three times a week. Subsequently, a mixed suspension of Caco-2 and HT29-MTX cells was seeded in DMEM + FBS at a density of $\approx 10^5$ cells/cm^2 into cell culture inserts with membrane filters (pore size 0.4 μm) for Falcon 24-well-multitrays (Becton, Dickinson & Company, Franklin Lakes, NJ, USA). Cells were grown for 21 days until a tight monolayer was formed (TEER > 600 Ω cm^2) with a medium replacement every three days.

Co-cultures of human cell monolayers and bifidobacterial

After 21 days from seeding, the culture medium of the 24-well plates was replaced with fresh, antibiotic-free DMEM. Subsequently, bifidobacterial cells with a final concentration of ≈10^8 cells/ml were inoculated on the Caco-2/HT29-MTX cell monolayers, as previously described[100]. The 24-well plates were then incubated at 5% CO_2 at 37°C. After 4h of incubation, bacterial cells were recovered in RNA later and stored at -80°C until processing.

For these experiments, *B. longum* subsp. *longum* PRL2022 (harboring the GH136-encoding gene) and *B. longum* subsp. *longum* 1898B (lacking the GH136-encoding gene) were grown in MRS broth in anaerobic conditions at 37°C. Once the exponential growth phase ($0.6 < OD_{600nm} < 0.8$) was reached, bifidobacterial cells were enumerated by using the Thoma cell counting chamber (Herka), diluted to reach a final concentration of 10^8 cells/ml, washed in PBS, resuspended in 400 µl of antibiotic-free DMEM, and seeded on Caco-2/HT29-MTX cell monolayers. Furthermore, bifidobacterial strains resuspended in DMEM and maintained at the same incubation conditions of the 24-well plates without any contact with human cell lines were used as sample control. All experiments were carried out in triplicate.

In addition, to test whether the hormonal fraction of the serum-supplement FBS did affect expression of the GH136 gene, *B. longum* subsp. *longum* PRL2022 was differentially grown on the culture medium DMEM with and without FBS, following the same protocol described above. Subsequently, the obtained RNA was used to perform qRT-PCR experiments to assess any differences in the expression of GH136 between the two considered conditions (DMEM+FBS or DMEM without FBS). Specifically, reverse transcription to cDNA was performed with the iScript Select cDNA synthesis kit (Bio-Rad Laboratories, USA) using the following thermal cycle: 5 min at 25°C, 20 min at 46°C, and 1 min at 95°C. The mRNA expression levels were

assessed with SYBR green technology in qRT-PCR using the Power Up SYBR Green Mastermix (ThermoFisher Scientific, USA) on a Bio-Rad CFX96 system according to the manufacturer's instructions. For this purpose,
rpoB-fw (CACGATGGTGCTGCGACCTTCCC),
rpoB-rv (GACCTGACGGATACGACGGTTGCC),
atpD-fw (CGTATGCCTTCCGCCGTGGGTTAC),
atpD-rv (ACGTAGATGGCTTGCAGCGAGGTG),
ldh-fw (GTGATGGGCGAGCATGGCGACTC),
ldh-rv (GGAGGCGAAGCGGTCTTGGTTGTC) were used as primers for the amplification of the housekeeping genes *rpoB*, *atpD*, and *ldh*, while primer pair GH136-fw (AGCGTCTCGAAGCACATCAA) and
GH136-rv (AGATCATCAGCGAGGCGAAG) was used to quantify GH136 gene expression. The PCR was carried out according to the following cycle: initial hold at 95°C

Eukaryotic RNA-Seq data analysis

Total RNA was extracted from human cell monolayers with RNeasy Mini Kit (Qiagen). All samples had an RNA integrity number (RIN) ≥ 8. For RNA sequencing (RNA-Seq), TruSeq Standard mRNA Sample preparation kit (Illumina, San Diego, USA) was used to prepare stranded cDNA libraries with poly dT enrichment from 0.1 µg to 4 µg of RNA extracted from each sample according to the manufacturer's instructions. The quality and quantity of each cDNA sample was assessed by a Tape station 2200 (Agilent Technologies, USA) and Qubit Fluorometer (Thermofisher). Subsequently, the cDNA libraries were sequenced using an Illumina NextSeq 500 high-output v2.5 kit (150 cycles, single end) (Illumina) according to the technical support guide. The fastq-MCF program was used for trimming RNA-Seq raw data (fastq) based on quality score and presence of adapter sequence. High quality fastq

were aligned to the Human reference genome sequence (GRCh38.p13) by using the splice-aware STAR algorithm (version 2.7.10a)[101], and the quality of alignments was evaluated using Picard software tool (version 2.26.11) (https://broadinstitute.github.io/picard/). Subsequently, quantification of reads mapped to individual gene transcripts was achieved through htseq-counts script of HTSeq software in "union" mode[95]. Raw counts were then normalized using CPM (Counts per million mapped reads) for filtering genes with low counts (CPM <1) and TMM (Trimmed Mean of M-Values) for statistically robust differential gene expression analysis through the EdgeR package[102]. The expression difference was evaluated as log2 fold change (logFC) of average expression in each sample pair of compared groups (co-cultures Caco-2/HT29-MTX/B. longum subsp *longum* 1898B and Caco-2/HT29-MTX/*B. longum* subsp. *longum* PRL2022). Additionally, a Volcano plot was created for each comparison to simultaneously visualize expression changes (log fold change) and their statistical significance (*p*-value).

B. longum subsp. *longum* RNA-Seq data analysis

Extraction of total RNA from *B. longum* subsp. *longum* strains, RNA sequencing, and raw fastq processing were performed as described above for the *B. bifidum* RNA-sequencing experiment. Generation of raw counts and identification of DEGs were achieved as described above for the eukaryotic RNA-seq data analysis.

Human retrospective clinical studies

Shotgun metagenomic datasets of fecal samples were retrieved from publicly available human clinical trials in which female and male subjects consumed daily doses of microbial formulations containing *B. longum* subsp. *longum* strains (PRJNA324129, PRJEB28097). The chromosomes of the bacterial strains used in the selected studies (ANI value of 98.5%) were manually inspected to assess the

presence/absence of the GH136 gene, resulting in two different treatment patterns. A group of 21 healthy volunteers (11 females and 10 males) received *B. longum* subsp. *longum* AH1206, which was found to possess the GH136 gene (daily dose of 10^{10} viable cells for a 4-week period). For this latter cohort, fecal samples were collected at the baseline period, after 28 days of treatment, and about 200 days after consumption cessation. A second group of healthy individuals (n = 10; 4 females and 6 males) consumed a probiotic supplement (5×10^9 CFU bi-daily for a 4-week period) containing *B. longum* subsp. *longum* devoid of the GH136 gene (named strain AG1). Corresponding fecal samples were analyzed at baseline, on day 21 of microbial intervention, and after 28 days of follow-up. Genome sequence of the strain AH1206 was acquired from the NCBI database, while the chromosome of AG1 was recovered from the publicly available metagenomic sequenced reads of the used probiotic supplement using Spades v3.15 (metagenomic mode) with default parameters. Taxonomic classification of the assembled contigs was achieved using METAnnotatorX2 pipeline with manually curated genome databases. Completeness and contamination of reconstructed chromosome *B. longum* subsp. *longum* AG1 were validated through CheckM analysis[87]. Test of *B. longum* subsp. *longum* persistence was performed through reads mapping with Bowtie2, as described above. Briefly, metagenomic reads were filtered to remove low-quality (score lower than 20), tRNA, and rRNA sequences. Moreover, the sequences mapping against *Homo sapiens* genome were also eliminated. Filtered reads were used as input for Bowtie2 (--very-sensitive option, with at least 99% full-length identity). HTSeq software (running in union mode) was used to calculate read counts corresponding to each reference genome[95]. For each metagenomic sample, the relative abundance of mapped reads was estimated by normalizing the raw reads count on the total number of sequenced reads.

Statistical analyses

The software SPSS version 25, and ORIGIN version 9.8.0.200 (www.ibm.com/software/it/analytics/spss/) (https://www.originlab.com/) were used for statistical data analyses and graphing. PERMANOVA was calculated using the adonis2 function from the vegan R package. PERMANOVA analyses based on Bray-Curtis measures of species-level abundance data were conducted using 1000 permutations to estimate *p*-values for the observed differences between the compared groups in PCoA analyses. In differential gene expression analysis, EdgeR package was used to estimate the statistical significance of differences between fold changes as the False Discovery Rate (FDR).

ACKNOWLEDGMENTS

We thank GenProbio Srl for the financial support of the Laboratory of Probiogenomics. Part of this research is conducted using the High-Performance Computing (HPC) facility of the University of Parma.

FUNDING ACKNOWLEDGEMENTS

This research has financially been supported by the Programme “FIL-Quota Incentivante” of University of Parma and co-sponsored by Fondazione Cariparma". DvS is a member of APC Microbiome Ireland funded by Science Foundation Ireland (SFI), through the Irish Government’s National Development Plan (Grant no. SFI/12/RC/2273-P1 and SFI/12/RC/2273-P2). G.T. has been supported by “Fondazione Cariparma” in the framework of the project entitled “Parma Microbiota”. LMV has been supported by by “Programma Operativo Nazionale 2014-2020 of the Italian Ministry of University and Research. The funding from Project AGL2017-83653R (Spanish “Ministerio de Ciencia, Innovación y

Universidades (MCIU)", "Agencia Estatal de Investigación (AEI)" and FEDER) is also acknowledged.

Contributions

Conceptualization, supervising, coordination, and revising of manuscript by F.T., O.V., D.v.S., C.M., and M.V.

RNAseq and bacterial DNA sequencing experiments by G.A., G.L., R.A., and A.V.

Collection data and genomic, metagenomic, and transcriptomic data processing and analysis by C.T., G.A., S.M.R., M.B., C.A., L.M.V., S.A., and M.G.

Assisting in computational framework development, implementation, and computational analyses by F.F., G.A.L., and L.M.

Drafting of manuscript by C.T., G.A., and C.M.

All authors critically revised and approved the manuscript.

Data availability. Raw sequences produced by shotgun metagenomic sequencing and RNA-seq are accessible through sequence read archive (SRA) accession number PRJNA833139.

Code availability. No custom code was used for this study.

References

1. Derrien, M., Alvarez, A. S. & de Vos, W. M. The Gut Microbiota in the First Decade of Life. Trends Microbiol 27, 997–1010 (2019).

2. O'Toole, P. W. & Jeffery, I. B. Gut microbiota and aging. Science 350, 1214–1215 (2015).

3. Valdes, A. M., Walter, J., Segal, E. & Spector, T. D. Role of the gut microbiota in nutrition and health. BMJ 361, 36–44 (2018).

4. Koenig, J. E. et al. Succession of microbial consortia in the developing infant gut microbiome. Proc Natl Acad Sci U S A 108 Suppl 1, 4578–4585 (2011).

5. Bäckhed, F. et al. Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe 17, 690–703 (2015).

6. Martín, V. et al. Sharing of bacterial strains between breast milk and infant feces. J Hum Lact 28, 36–44 (2012).

7. Milani, C. et al. The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. Microbiology and Molecular Biology Reviews 81, (2017).

8. Duranti, S. et al. Maternal inheritance of bifidobacterial communities and bifidophages in infants through vertical transmission. Microbiome 5, (2017).

9. Korpela, K. et al. Selective maternal seeding and environment shape the human gut microbiome. Genome Res 28, 561–568 (2018).

10. Pannaraj, P. S. et al. Association Between Breast Milk Bacterial Communities and Establishment and Development of the Infant Gut Microbiome. JAMA Pediatr 171, 647–654 (2017).

11. Bäckhed, F. et al. Erratum: Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life (Cell Host and Microbe (2015) 17(5) (690–703)). Cell Host Microbe 17, 852 (2015).

12. Jeurink, P. V. et al. Human milk: a source of more life than we imagine. Benef Microbes 4, 17–30 (2013).

13. Scanlan, P. D. Microbial evolution and ecological opportunity in the gut environment. Proc Biol Sci 286, (2019).

14. Marshall, C. G., Ogden, D. C. & Colquhoun, D. The actions of suxamethonium (succinyldicholine) as an agonist and channel blocker at the nicotinic receptor of frog muscle. J Physiol 428, 155–174 (1990).

15. Linehan, K., Dempsey, E. M., Ryan, C. A., Ross, R. P. & Stanton, C. First encounters of the microbial kind: perinatal factors direct infant gut microbiome establishment. Microbiome Research Reports 1, 10 (2022).

16. Sangwan, N., Xia, F. & Gilbert, J. A. Recovering complete and draft population genomes from metagenome datasets. Microbiome 4, (2016).

17. Lugli, G. A. et al. Genetic insights into the dark matter of the mammalian gut microbiota through targeted genome reconstruction. Environ Microbiol 23, 3294–3305 (2021).

18. Lugli, G. A. & Ventura, M. A breath of fresh air in microbiome science: shallow shotgun metagenomics for a reliable disentangling of microbial ecosystems. Microbiome Research Reports 1, 8 (2022).

19. Dizzell, S. et al. Investigating colonization patterns of the infant gut microbiome during the introduction of solid food and weaning from breastmilk: A cohort study protocol. PLoS One 16, (2021).

20. Coker, M. O. et al. Infant Feeding Alters the Longitudinal Impact of Birth Mode on the Development of the Gut Microbiota in the First Year of Life. Front Microbiol 12, (2021).

21. Odamaki, T. et al. Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. BMC Microbiol 16, (2016).

22. Milani, C. et al. Exploring Vertical Transmission of Bifidobacteria from Mother to Child. Appl Environ Microbiol 81, 7078–7087 (2015).

23. Nilsen, M. et al. Butyrate Levels in the Transition from an Infant- to an Adult-Like Gut Microbiota Correlate with Bacterial Networks Associated with Eubacterium Rectale and Ruminococcus Gnavus. Genes (Basel) 11, 1–15 (2020).

24. Moore, R. E. & Townsend, S. D. Temporal development of the infant gut microbiome. Open Biol 9, 190128 (2019).

25. Hildebrand, F. et al. Dispersal strategies shape persistence and evolution of human gut bacteria. Cell Host Microbe 29, 1167 (2021).

26. Sela, D. A. Bifidobacterial utilization of human milk oligosaccharides. Int J Food Microbiol 149, 58–64 (2011).

27. Turroni, F. et al. Deciphering bifidobacterial-mediated metabolic interactions and their impact on gut microbiota by a multi-omics approach. ISME J 10, 1656–1668 (2016).

28. Turroni, F. et al. Bifidobacteria and the infant gut: an example of co-evolution and natural selection. Cell Mol Life Sci 75, 103–118 (2018).

29. Shao, Y. et al. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. Nature 574, 117–121 (2019).

30. Mäklin, T. et al. Strong pathogen competition in neonatal gut colonisation. Nature Communications 2022 13:1 13, 1–13 (2022).

31. Sakanaka, M. et al. Varied Pathways of Infant Gut-Associated Bifidobacterium to Assimilate Human Milk Oligosaccharides: Prevalence of the Gene Set and Its Correlation with Bifidobacteria-Rich Microbiota Formation. Nutrients 12, (2019).

32. Murphy, R. et al. Eczema-protective probiotic alters infant gut microbiome functional capacity but not composition: sub-sample analysis from a RCT. Benef Microbes 10, 5–17 (2019).

33. Olm, M. R. et al. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nature Biotechnology 2021 39:6 39, 727–736 (2021).

34. Moubareck, C. A. Human Milk Microbiota and Oligosaccharides: A Glimpse into Benefits, Diversity, and Correlations. Nutrients 13, (2021).

35. Tarracchini, C. et al. Phylogenomic disentangling of the bifidobacterium longum subsp. infantis taxon. Microb Genom 7, (2021).

36. Yamada, C. et al. Molecular Insight into Evolution of Symbiosis between Breast-Fed Infants and a Member of the Human Gut Microbiome Bifidobacterium longum. Cell Chem Biol 24, 515-524.e5 (2017).

37. Yoon, K. & Kim, N. Roles of Sex Hormones and Gender in the Gut Microbiota. J Neurogastroenterol Motil 27, 314 (2021).

38. Valeri, F. & Endres, K. How biological sex of the host shapes its gut microbiota. Front Neuroendocrinol 61, 100912 (2021).

39. Özkurt, E. & Hildebrand, F. Lifelong sex-dependent trajectories of the human gut microbiota. Nature Aging 2021 1:1 1, 22–23 (2021).

40. Kim, Y. S., Unno, T., Kim, B. Y. & Park, M. S. Sex Differences in Gut Microbiota. World J Mens Health 38, 48 (2020).

41. Gomez, A., Luckey, D. & Taneja, V. The gut microbiome in autoimmunity: Sex matters. Clin Immunol 159, 154–162 (2015).

42. Sisk-Hackworth, L., Kelley, S. T. & Thackray, V. G. Sex, puberty, and the gut microbiome. Reproduction 165, R61–R74 (2023).

43. Croix, J. A. et al. On the relationship between sialomucin and sulfomucin expression and hydrogenotrophic microbes in the human colonic mucosa. PLoS One 6, (2011).

44. Diebel, M. E., Diebel, L. N., Manke, C. W. & Liberati, D. M. Estrogen modulates intestinal mucus physiochemical properties and protects against oxidant injury. Journal of Trauma and Acute Care Surgery 78, 94–99 (2015).

45. Choi, H. J. et al. Signal pathway of 17beta-estradiol-induced MUC5B expression in human airway epithelial cells. Am J Respir Cell Mol Biol 40, 168–178 (2009).

46. Gollub, E. G., Waksman, H., Goswami, S. & Marom, Z. Mucin genes are regulated by estrogen and dexamethasone. Biochem Biophys Res Commun 217, 1006–1014 (1995).

47. Courant, F. et al. Assessment of circulating sex steroid levels in prepubertal and pubertal boys and girls by a novel ultrasensitive gas chromatography-tandem mass spectrometry method. Journal of Clinical Endocrinology and Metabolism 95, 82–92 (2010).

48. Kuiri-Hänninen, T., Sankilampi, U. & Dunkel, L. Activation of the Hypothalamic-Pituitary-Gonadal Axis in Infancy: Minipuberty. Horm Res Paediatr 82, 73–80 (2014).

49. Becker, M. & Hesse, V. Minipuberty: Why Does it Happen? Horm Res Paediatr 93, 76–84 (2020).

50. [Sex differences in the secretion of gonadotropins and sex hormones in newborns and infants] - PubMed. https://pubmed.ncbi.nlm.nih.gov/6767647/.

51. Becker, M. et al. Hormonal ‘minipuberty’ influences the somatic development of boys but not of girls up to the age of 6 years. Clin Endocrinol (Oxf) 83, 694–701 (2015).

52. Kuiri-Hänninen, T., Dunkel, L. & Sankilampi, U. Sexual dimorphism in postnatal gonadotrophin levels in infancy reflects diverse maturation of the ovarian and testicular hormone synthesis. Clin Endocrinol (Oxf) 89, 85–92 (2018).

53. Kaz, A. M. et al. Patterns of DNA methylation in the normal colon vary by anatomical location, gender, and age. Epigenetics 9, 492–502 (2014).

54. Pinho, R. M. & Maga, E. A. DNA methylation as a regulator of intestinal gene expression. Br J Nutr 126, 1611–1625 (2021).

55. Yousefi, P. et al. Sex differences in DNA methylation assessed by 450K BeadChip in newborns. BMC Genomics 16, 1–12 (2015).

56. Duranti, S. et al. Bifidobacterium bifidum and the infant gut microbiota: an intriguing case of microbe-host co-evolution. Environ Microbiol 21, 3683–3695 (2019).

57. Rodríguez, J. M. The origin of human milk bacteria: is there a bacterial entero-mammary pathway during late pregnancy and lactation? Adv Nutr 5, 779–784 (2014).

58. Ferretti, P. et al. Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host Microbe 24, 133-145.e5 (2018).

59. Makino, H. et al. Transmission of intestinal Bifidobacterium longum subsp. longum strains from mother to infant, determined by multilocus sequencing typing and amplified fragment length polymorphism. Appl Environ Microbiol 77, 6788–6793 (2011).

60. Solís, G., de los Reyes-Gavilan, C. G., Fernández, N., Margolles, A. & Gueimonde, M. Establishment and development of lactic acid bacteria and bifidobacteria microbiota in breast-milk and the infant gut. Anaerobe 16, 307–310 (2010).

61. Yassour, M. et al. Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. Cell Host Microbe 24, 146-154.e4 (2018).

62. Bäckhed, F. et al. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe 17, 690–703 (2015).

63. Maldonado-Gómez, M. X. et al. Stable Engraftment of Bifidobacterium longum AH1206 in the Human Gut Depends on Individualized Features of the Resident Microbiome. Cell Host Microbe 20, 515–526 (2016).

64. Zmora, N. et al. Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated with Unique Host and Microbiome Features. Cell 174, 1388-1405.e21 (2018).

65. Sperandio, V., Torres, A. G., Jarvis, B., Nataro, J. P. & Kaper, J. B. Bacteria-host communication: the language of hormones. Proc Natl Acad Sci U S A 100, 8951–8956 (2003).

66. Karavolos, M. H. & Anjam Khan, C. M. Multidirectional chemical signalling between Mammalian hosts, resident microbiota, and invasive pathogens: neuroendocrine hormone-induced changes in bacterial gene expression. Adv Exp Med Biol 817, 241–253 (2014).

67. García-Gómez, E., González-Pedrajo, B. & Camacho-Arroyo, I. Role of sex steroid hormones in bacterial-host interactions. Biomed Res Int 2013, (2013).

68. Kaisho, T. & Akira, S. Toll-like receptors as adjuvant receptors. Biochim Biophys Acta 1589, 1–13 (2002).

69. Fukata, M. & Arditi, M. The role of pattern recognition receptors in intestinal inflammation. Mucosal Immunol 6, 451–463 (2013).

70. Hosoi, T. et al. Cytokine responses of human intestinal epithelial-like Caco-2 cells to the nonpathogenic bacterium Bacillus subtilis (natto). Int J Food Microbiol 82, 255–264 (2003).

71. Melhem, H., Regan-Komito, D. & Niess, J. H. Mucins Dynamics in Physiological and Pathological Conditions. Int J Mol Sci 22, (2021).

72. González-Vallinas, M. et al. Clinical relevance of the differential expression of the glycosyltransferase gene GCNT3 in colon cancer. Eur J Cancer 51, 1–8 (2015).

73. Tan, S. & Cheng, P. W. Mucin biosynthesis: identification of the cis-regulatory elements of human C2GnT-M gene. Am J Respir Cell Mol Biol 36, 737–745 (2007).

74. Wang, S. et al. Metagenomic analysis of mother-infant gut microbiome reveals global distinct and shared microbial signatures. Gut Microbes 13, 1–24 (2021).

75. Mäklin, T. et al. High-resolution sweep metagenomics using fast probabilistic inference. Wellcome Open Res 5, 14 (2021).

76. Mäklin, T. et al. Bacterial genomic epidemiology with mixed samples. Microb Genom 7, (2021).

77. Alanko, J. N., Vuohtoniemi, J., Mäklin, T. & Puglisi, S. J. Themisto: a scalable colored k-mer index for sensitive pseudoalignment against hundreds of thousands of bacterial genomes. bioRxiv 2023.02.24.529942 (2023) doi:10.1101/2023.02.24.529942.

78. Milani, C. et al. METAnnotatorX2: a Comprehensive Tool for Deep and Shallow Metagenomic Data Set Analyses. mSystems 6, (2021).

79. Chen, Y., Ye, W., Zhang, Y. & Xu, Y. High speed BLASTN: An accelerated MegaBLAST search tool. Nucleic Acids Res 43, 7762–7768 (2015).

80. Bankevich, A. et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19, 455–477 (2012).

81. Lugli, G. A. et al. Isolation of novel gut bifidobacteria using a combination of metagenomic and cultivation approaches. Genome Biol 20, (2019).

82. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, (2010).

83. Lugli, G. A., Milani, C., Mancabelli, L., Van Sinderen, D. & Ventura, M. MEGAnnotator: a user-friendly pipeline for microbial genomes assembly and annotation. FEMS Microbiol Lett 363, (2016).

84. Lugli, G. A. et al. MEGAnnotator2: a pipeline for the assembly and annotation of microbial genomes. Microbiome Research Reports 2, 15 (2023).

85. van Dijk, L. R. et al. StrainGE: a toolkit to track and characterize low-abundance strains in complex microbial communities. Genome Biol 23, (2022).

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

87. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25, 1043–1055 (2015).

88. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42, (2014).

89. Zhao, Y. et al. PGAP: pan-genomes analysis pipeline. Bioinformatics 28, 416–418 (2012).

90. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30, 1575–1584 (2002).

91. Letunic, I., Khedkar, S. & Bork, P. SMART: recent updates, new developments and status in 2020. Nucleic Acids Res 49, D458–D460 (2021).

92. Waterhouse, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 46, W296–W303 (2018).

93. Tarracchini, C. et al. Assessing the Genomic Variability of Gardnerella vaginalis through Comparative Genomic Analyses: Evolutionary and Ecological Implications. Appl Environ Microbiol 87, 1–16 (2020).

94. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359 (2012).

95. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166 (2015).

96. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589–595 (2010).

97. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, (2014).

98. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5, 621–628 (2008).

99. Bianchi, M. G. et al. Catechin and Procyanidin B 2 Modulate the Expression of Tight Junction Proteins but Do Not Protect from Inflammation-Induced Changes in Permeability in Human Intestinal Cell Monolayers. Nutrients 11, (2019).

100. Serafini, F. et al. Evaluation of adhesion properties and antibacterial activities of the infant gut commensal Bifidobacterium bifidum PRL2010. Anaerobe 21, 9–17 (2013).

101. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).

102. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140 (2010).

Chapter 4

Phylogenomic disentangling of the *Bifidobacterium longum* subsp. *infantis* taxon

Chiara Tarracchini, Christian Milani, Gabriele Andrea Lugli, Leonardo Mancabelli, Federico Fontana, Giulia Alessandri, Giulia Longhi, Rosaria Anzalone, Alice Viappiani, Francesca Turroni, Douwe van Sinderen, Marco Ventura.

The results of this chapter were published in Microbial Genomics, 2021 Jul; doi: 10.1099/mgen.0.000609.

Abstract

Members of the *Bifidobacterium longum* species have been shown to possess adaptive abilities to allow colonization of different mammalian hosts, including humans, primates and domesticated mammalian species, such as dogs, horses, cattle and pigs. To date, three subspecies have formally been recognized to belong to this bifidobacterial taxon, i.e., *B. longum* subsp. *longum*, *B. longum* subsp. *infantis* and *B. longum* subsp. *suis*. Although *B. longum* subsp. *longum* is widely distributed in the human gut irrespective of host age, *B. longum* subsp. *infantis* appears to play a significant role as a prominent member of the gut microbiota of breast-fed infants. Nevertheless, despite the considerable scientific relevance of these taxa and the vast body of genomic data now available, an accurate dissection of the genetic features that comprehensively characterize the *B. longum* species and its subspecies is still missing. In the current study, we employed 261 publicly available *B. longum* genome sequences, combined with those of 11 new isolates, to investigate genomic diversity of this taxon through comparative genomic and phylogenomic approaches. These analyses allowed us to highlight a remarkable intra-species genetic and physiological diversity. Notably, the more extensive genome content observed for members of the *B. longum* subsp. *infantis* subspecies was shown to be linked to the acquisition of genetic features that appear to endow it with increased competitiveness in the gut environment of suckling hosts. Furthermore, specific *B. longum* subsp. *infantis* genomic features appear to be responsible for enhanced Horizontal Gene Transfer (HGT) occurrences, underpinning a remarkable dedication toward acquisition of foreign DNA by HGT events, thereby earning itself the title of “genetic vacuum cleaner” of the gut microbiota.

Impact statement

In this study, through comparative genomic analyses and phylogenomic reconstruction of 261 publicly available *B. longum* genomes, we gained insight into intra-species genetic and physiological diversity, identifying specific *B. longum* subsp. *infantis* genomic features which appear to be linked with its enhanced ability to acquire foreign DNA. This remarkable genome plasticity may provide an explanation for the specific adaptation of *B. longum* subsp. *infantis* toward colonization of the gut of suckling mammals.

Data summary

Decoded genome sequences of 11 newly isolated *B. longum* strains were deposited at NCBI database under BioProject code PRJNA692178. A full listing of NCBI accession data for *B. longum* strains described in this paper is available in Table S1.

For Supplementary Materials see the article published in Microbial Genomics

INTRODUCTION

The human gut harbors at least 100 trillion (10^{14}) microbial cells [1], collectively organized in a complex and dynamic microbial community that plays a fundamental role in defining human health status [2]. It is well known that members of the gut microbiota engage in complex microbe-microbe and microbe-host interactions, with physiological consequences, including participation in metabolic activities such as (sometimes syntrophic) degradation of non-digestible carbohydrates, with consequent production of short-chain fatty acids (SCFAs) [3, 4]. The assembly of the human gut microbiota is believed to commence during delivery when the newborn passes through the mother's birth canal [5]. During the developmental period following birth, the early gut microbiota is influenced by various factors, including mode of delivery, duration of gestation, antibiotic exposure, as well as feeding type [6, 7]. This latter factor is particularly noteworthy since breast-feeding can shape the gut microbiota composition of the newborn by promoting a microbial community enriched by members of the *Bifidobacterium* genus [8]. In addition to the fermentation of non-digestible food compounds, in particular glycans, the (bifido)bacterial consortia also engage with the host immune system, stimulating and modulating both innate and adaptive host immune responses, ultimately influencing overall intestinal functionality and homeostasis [9, 10]. Interestingly, it has been reported that particular bifidobacterial species, such as *Bifidobacterium longum* subsp. *infantis*, *Bifidobacterium bifidum,* and *Bifidobacterium breve*, are able to efficiently utilize (certain) human milk oligosaccharides (HMOs) [11-15]. HMOs constitute complex milk glycans known to elicit prebiotic activity by allowing the above-mentioned bifidobacterial species to establish and persist in the infant gut, thereby representing a clear example of host-microbe co-evolution in humans [16-20].

Members of the *Bifidobacterium longum* species have been identified as very common inhabitants of the mammalian gut, reaching a prevalence of 95.5 %, representing the percentage of individuals harboring this species within the population, as shown by a recent survey conducted in 67 assessed mammalian hosts [21]. In recent decades, members of the *B. longum* species have been grouped into three distinct subspecies**,** i.e., *Bifidobacterium longum* subsp. *longum*, *Bifidobacterium longum* subsp. *infantis* and *Bifidobacterium longum* subsp. *suis* [22], the latter isolated from the gut microbiota of swine [22, 23]. Despite the progressive reduction in the relative abundance of bifidobacteria in the human gut starting from one/two years of age [7], members of *B. longum* subsp. *longum* are known to commonly inhabit the infant, adult and elderly human gut [24], perhaps exerting their positive health footprint throughout the human lifespan [24, 25]. In contrast, *B. longum* subsp. *infantis* is most frequently isolated from breast-fed infant feces [26, 27]. Consistently, the decoding of *B. longum* subsp. *infantis* ATCC15697 genome sequence, which was published in 2008, revealed a genome that is dedicated to the degradation and utilization of a wide range of HMOs [15, 28].

Due to the substantial scientific and commercial interest in members of this species, which are able to colonize different hosts at different stages of life, during which they may contribute to host health, a large number of *B. longum* strains have been sequenced. Nevertheless, a comprehensive dissection of the genetic potential of *B. longum* and its subspecies is still lacking. For this reason, we decided to investigate the genomic diversity of and phylogenetic relationships between members of the *B. longum* species. This prompted a complete revision of subspecies classification and allowed a detailed dissection of their genetic features presumed to be responsible for efficient niche adaptation.

RESULTS AND DISCUSSION

General genome features of *B. longum* genomes included in the comparative genomics analysis

In order to investigate the phylogenomic diversity of members belonging to the *B. longum* species, we undertook a comparative genomics analysis involving high-quality *B. longum* genome sequences selected amongst those publicly available (complete and draft genome sequences, see M&M section for the inclusion/exclusion criteria used). Remarkably, among the latter, *B. longum* subsp. *infantis* strains exhibited the highest number of suspected duplicated genomes (ANI ≥ 99.99 %). Accordingly, we removed such apparent copies of identical chromosomes that had been deposited under different strain IDs, thereby allowing the generation of a curated *B. longum* subsp. *infantis* genome collection without duplicated chromosomal sequences (Table S1).

The final collection of 272 *B. longum* genomes, including the 11 sequenced in this study, encompassed chromosomal sequences ranging in size from 2.2 Mb for *B. longum* APC1478 to 2.8 Mb for *B. longum* subsp. *infantis* ATCC 15697. As outlined in Table S1, the number of predicted Coding DNA Sequences (CDS) ranged from 1,685 for *B. longum* subsp. *longum* 296B to 2,412 for *B. longum* subsp. *infantis* ATCC 15697, with an average value of 1,927.17 ± 114.61 CDSs per genome (Table S1). Notably, the chromosomes belonging to the *B. longum* subsp. *infantis* subspecies emerged as the largest ones among the assessed *B. longum* genomes, ranging in size between 2.6 Mb and 2.8 Mb (ANOVA p-value < 0.05). These results showed that genome size might vary considerably even in closely related strains of the same species, thus indicating remarkable intra-species genetic and physiological diversity, unlike what was previously found for other bifidobacterial species such as *Bifidobacterium bifidum* and *Bifidobacterium dentium* [46, 47].

Pan-genome and core genome of *B. longum* species

In recent years, computation of the pan-genome has been employed as an approach to investigate overall genomic differences and infer precise phylogenomic relationships between (bifido)bacterial taxa [29, 48-51]. Accordingly, the genomes of *B. longum* strains were subjected to pan-genome analysis, allowing the identification of a total of 22,591 Clusters of Orthologous Groups (COGs). Analysis of the pan-genome size increasing as genomes are sequentially included showed an average of 49.7 newly added COGs at the last three iterations (see Supplementary text for details). This trend is indicative of a pan-genome that has not yet fully reached its completion, though approaching a saturation plateau (Figure S1).

A total of 510 COGs were classified as a collection of genes shared by all assessed strains, thereby representing the core genome of the *B. longum* species. Furthermore, the Truly Unique Genes (TUGs) for each *B. longum* strain were also identified, revealing an average of 48.5 TUGs per genome (see Supplementary text for details). The relatively small number of core genes observed suggests the presence of rather high intra-species variability, particularly when compared to other previously investigated bifidobacterial species, such as *B. bifidum*, *Bifidobacterium breve* (1,295 and 1,307 conserved COGs, respectively) [46, 52]. On the other hand, the relatively small number of TUGs is comparable with that previously observed for the genomes of *Bifidobacterium pseudolongum* and *B. dentium* (41 and 60 average TUGs, respectively) [29, 47], implying that a large part of the genetic diversity resides in the dispensable gene pool, i.e., those genes that are shared by a subgroup of strains, possibly due to adaptation to specific ecological niches/hosts.

Interestingly, *B. pseudolongum* species, for which the subspecies *pseudolongum* and *globosum* are recognized, showed a much larger number of core genes, i.e., 1,069 COGs, when compared to those identified in *B. longum* genomes. Therefore, these findings suggest that the latter taxon is characterized by a relatively high intra-

specific variability, which may be imputed to distinct genetic traits possessed by each *B. longum* subspecies.

Phylogenetic analyses the *B. longum* taxon

The pairwise percentage Average Nucleotide Identity (ANI) is currently considered to represent the gold standard for inference of close phylogenetic relationships and (sub)species classification of bacterial genomes [40]. Evaluation of the overall genomic differences between the 271 *B. longum* genomes through ANI analysis resulted in values ranging from 94.2 % to 98.9 % (Table S2). Notably, previous *Bifidobacterium* phylogenomic studies showed that an ANI threshold value of 94 % properly discriminates between bifidobacterial species [51, 53], being consistent with what has been observed for other phylogenetically related taxonomic groups in the Bifidobacteriaceae family, such as *Gardnerella* [54]. Accordingly, the finding that this phylogenomic analysis generated ANI values above 94.2 % indicates that the included genome sequences correctly fall within the boundaries of a single species, i.e., *B. longum*. Nonetheless, based on the ANI matrix (Table S2), it was possible to identify three subgroups corresponding to the three so far recognized subspecies of *B. longum*, within which the observed ANI values ranged from 96.3 % to 98.9 % (Table S2).

In order to precisely track the phylogenetic relationships between the strains of this species, we computed a phylogenetic tree based on the amino acid sequence alignment of the 510 COGs that constitute the core genome of this species (Figure S2). Due to the high number of analyzed genomes belonging to the *B. longum* subsp. *longum* subspecies, we decided to generate an additional tree encompassing a pool of 42 representative genomes of this taxon, chosen to maximize the genetic diversity

coverage, in order to obtain a clearer graphical visualization of the complete *B. longum* phylogeny (Figure 1) (See Supplementary text for details).

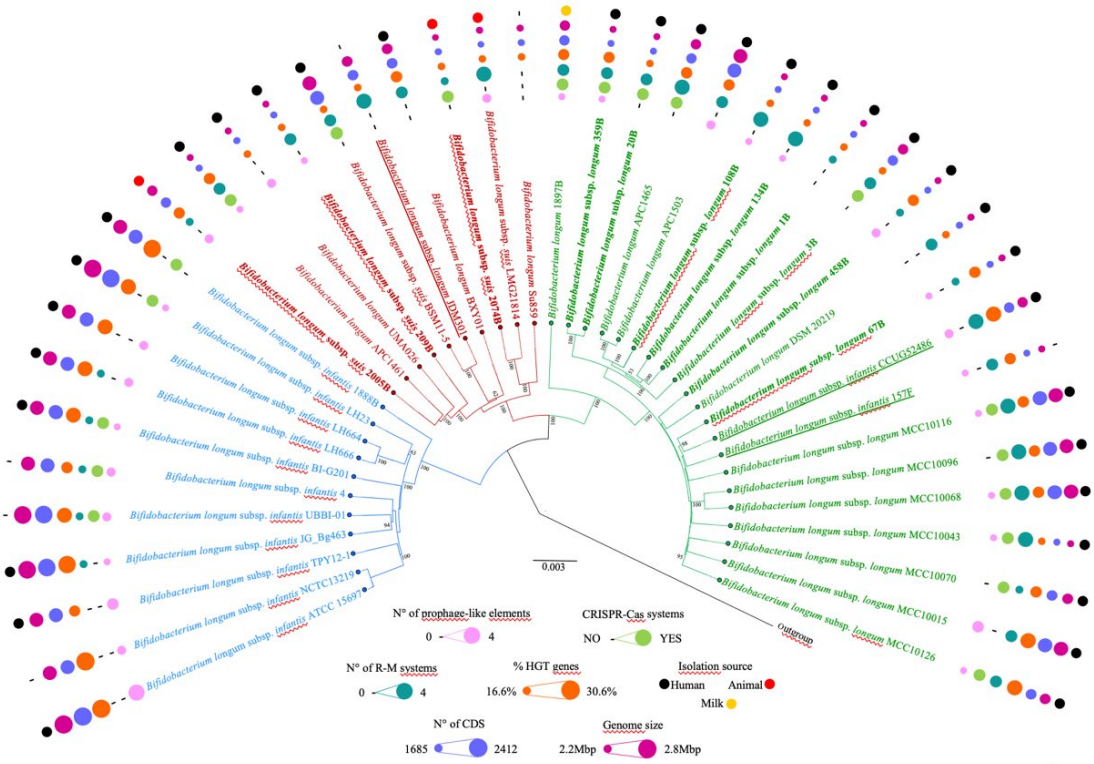


Figure 1. Phylogenomic tree based on the core genome of *B. longum* species. The phylogenomic tree, showing a selection of 42 representative genomes belonging to the *B. longum* species, was based on the concatenation of the 510 *B. longum* core genes and was built through the neighbor-joining method. Bootstrap percentages above 50 are shown at node points, based on 1,000 replicates. Misclassified strains were underlined while the 11 new isolates were highlighted in bold. Phylogenetic clusters are highlighted with similarly colored branches. Circles surrounding the tree represent the genome sizes (in dark pink), numbers of CDS (in purple), percentages of genes predicted to have undergone horizontal gene transfer (in orange), number of R-M systems (in dark green), occurrence of CRISPR-Cas systems (in light green), and isolation source (black=human, red=animal, yellow=milk).

As expected, the resulting *B. longum*-based phylogenetic tree revealed the presence of three main clades (Figure S2; Figure 1), consisting of the *B. longum* subsp. *longum* taxonomic group (*Bll*), the *B. longum* subsp. *infantis* taxonomic group (*Bli*) and the *B. longum* subsp. *suis* (*Bls*) taxonomic group (Figure 1).

In depth analysis of the tree revealed that strains *B. longum* subsp. *infantis* 157F [55], *B. longum* subsp. *infantis* CCUG 52486 [56] and *B. longum* subsp. *longum* JDM301 [57] had been misclassified. Specifically, consistent with what had previously been observed through ANI analysis (Table S2), strains 157F and CCUG 52486 had been assigned to the *B. longum* subsp. *longum* subspecies, while JDM301 had been classified as a member of the *B. longum* subsp. *suis* subspecies. Interpretation of the phylogenomic tree suggests a clear phylogenetic separation between members of *B. longum* subsp. *infantis* cluster and the other *B. longum* strains, indicative of earlier speciation with respect to *B. longum* subsp. *longum* and *B. longum* subsp. *suis,* which showed a closer phylogenetic relationship (Figure S2; Figure 1; Table S2).

Moreover, the phylogenomic-based approach, combined with ANI value assignment, was applied to taxonomically classify the 11 newly isolated *B. longum* strains in other to include them in subspecies-specific analyses (see below). Specifically, three genomes were shown to belong to *B. longum* subsp. *suis* subspecies, i.e., 209B, 2015B and 2074B, while the remaining eight were classified as members *B. longum* subsp. *longum* subspecies (Figure 1). Interestingly, *B. longum* subsp. *longum* 1897B, which had been isolated from human milk, was shown to belong to a separate branch with respect to all other *B. longum* subsp. *longum* strains (Figure 1), denoting a different evolutionary history compared to the other assessed *B. longum* members isolated from the mammalian gut.

The Pan- and Core- genome of the *B. longum* subspecies. Evolutionary processes have shaped bacterial genomes by driving changes in their genetic repertoire in order to facilitate adaptation to a specific environmental niche [58, 59], thus leading to (sub)speciation events. Pan-genome reconstruction may provide insights into these evolutionary events by unveiling genomic peculiarities and shared genetic traits that characterize a given bacterial taxon [60]. In the context of a *B. longum* subspecies-

focused comparative analysis, we separately analyzed subspecies-specific pan-genomes (Figure S3) (See Supplementary text for details). The 251 *B. longum* subsp. *longum* genomes and the ten members of *B. longum* subsp. *suis* used in these analyses showed similar average genome sizes, i.e., of 2.39 Mb and 2.43 Mb (Table S1). These latter are significantly smaller compared to that observed for *B. longum* subsp. *infantis* (average of 2.65 Mb) (Table S1), which also showed an average of 253 additional CDSs when compared to those found in *B. longum* subsp. *longum* and *B. longum* subsp. *suis* genomes (ANOVA p-value < 0.001) (Details in Supplementary text). This finding suggests that members of the *B. longum* subsp. *infantis* taxon may have evolved as a result of progressive acquisition of new genetic features [58].

The subspecies-specific pan-genome analyses also allowed the definition of the *Bll*-, *Bls*- and *Bli*-Core Genome (CG), intended as the subspecies-specific core genes repertoire. In detail, these subspecies-specific core genomes were defined by taking into account those COGs shared by at least 85 % of the strains belonging to a given *B. longum* subspecies while being absent in the other two subspecies. The decision to consider an 85 % gene sharing level, rather than the typically employed 100 %, was motivated by the presence of a high number of draft genomes within the analyzed genome collection, which therefore could influence the accuracy of the calculation of subspecies-specific core genomes. In this manner, a total of 24 and five core genes represented the *Bll*-CG and *Bls*-CG, respectively, whereas 53 genes were identified as constituting *Bli*-CG (Figure S3) (Details are reported in Supplementary text).
The relatively small size of the *Bll*-CG and *Bls*-CG may, at least in part, be due to their close phylogenetic relationship and to the high number of analyzed *Bll* genomes. However, it suggests that the evolutionary path taken by these subspecies may not have led to the acquisition of a substantial number of subspecies-specific competencies compared to their common *B. longum* ancestor. In contrast, the higher

number of genes constituting *Bli*-CG suggests that this subspecies was subject to a higher evolutionary pressure that instigated the acquisition of novel genetic traits. Interestingly, 31 (58 %) of *Bli*-CG, 17 (71 %) of *Bll*-CG and four of the five (80 %) of *Bls*-CG were found in other bifidobacterial species with identity > 50 % and coverage > 80 % by BLASTp search in currently available bifidobacterial genomes (Table S3). These data suggest, at first glance, that a subgroup of subspecies-specific core genes may have been acquired by a common bifidobacterial ancestor (as indicated by presence in other bifidobacteria) and subsequently lost at subspecies level. Nevertheless, each subspecies seems to have independently acquired new genetic features, with *B. longum* subsp. *infantis* showing the highest number of genes acquired by presumed HGT events (18.8 % of the *Bli*-CG) (Table 1).

Table 1. *B. longum* subspecies-specific core genes.

B. longum* subsp. *longum

Core Gene	Prevalence across the subspecies	Function		Transporter Classification Database		HGT events
		Interpro Database	Refseq Database	Function	Family	
B1_0665	99%	Selenoprotein, putative	YbdD/YjiX family protein			Native
B1_0666	98%	5TM C-terminal transporter carbon starvation CstA	Carbon starvation protein A	Peptide Transporter Carbon Starvation CstA (CstA) Family	2.A.114.-	Native
B1_1343	98%	Protein of unknown function (DUF3073)	DUF3073 domain-containing protein			Native
B1_0094	98%	NADH Oxidase	nitroreductase			Native
B1_0106	98%	Periplasmic binding protein-like II	extracellular solute-binding protein			Native
B1_0884	98%	-	aldo/keto reductase family protein			Native
B1_1277	97%	Glycosidases	Pullulanase type I			Native
B1_0628	97%	L,D-transpeptidase YCIB-related	L,D-transpeptidase			Native
B1_0345	97%	Metal-dependent hydrolase	Amidohydrolase family protein			Native
B1_1278	97%		Alpha-amylase			Native
B1_0156	95%	-	DUF2400 domain-containing protein			Native
B1_1275	95%	ABC transporter permease protein MG189-related	ABC transporter permease subunit	It binds α-(1,6)-linked glucosides and galactosides	3.A.1.1.53	Native
B1_0738	94%	-	DUF1846 domain-containing protein			Native
B1_1795	93%	Acyl-CoA N-acyltransferases (Nat)	GNAT family N-acetyltransferase			Native

B1_0431	91%	Uncharacterized protein conserved in bacteria C-term(DUF2220)	DUF3322 and DUF2220 domain-containing protein			Native
B1_1294	90%	-	Substrate-binding domain-containing protein			Native
B1_0735	90%	-	DUF87 domain-containing protein			Foreign
B1_0883	90%	Transcriptional dual regulator hcar-related	LysR family transcriptional regulator			Native
B1_0737	89%	Type VII secretion system protein EsaG-like	-			Foreign
134B_0607	88%	-	DNA/RNA non-specific endonuclease			Native
134B_0472	88%	MATE_MepA_like	MATE family efflux transporter			Native
B1_1296	88%	K+ potassium transporter	KUP/HAK/KT family potassium transporter			Native
B1_0107	86%	Glycosidase family 31	Alpha-xylosidase			Native
B1_0786	86%	zinc-ribbon domain	Zinc ribbon domain-containing protein			Native

B. longum* subsp. *infantis

Core Gene	Prevalence across the subspecies	Function		Transporter Classification Database		HGT events
		Interpro Database	Refseq Database	Function	Family	
ACJ51545.1	100%	MFS general substrate transporter domains	MFS transporter	Glucose Transporter (GT) Family	2.A.1.68.1	Native
ACJ53225.1	100%	Tetratricopeptide-like helical domain	DUF4037 domain-containing protein			Native
ACJ52470.1	100%	Ttransporter solute:sodium symporter family	Sodium/solute symporter	Glucose or galactose:Na+ symporter	2.A.1.68.1	Native
ACJ52099.1	100%	Response regulator receiver domain	Response regulator transcription factor			Foreign
ACJ51227.1	100%	-	-			Native
ACJ52098.1	100%	Lantibiotic immunity protein Spa1	NisI/SpaI family lantibiotic immunity			Native
ACJ53071.1	100%	Pyridoxal-phosphate dependent enzyme	Pyridoxal-phosphate dependent enzyme			Native
ACJ51238.1	100%	High-affinity nickel-transport protein	nickel/cobalt transporter			Native
ACJ51551.1	100%	Bacteriocin (Lactococcin_972)	Lactococcin 972 family bacteriocin			Foreign
ACJ51549.1	100%	Nitrate/nitrite sensor protein narx-related	Histidine kinase			Native
ACJ53072.1	100%	metallo-dependent hydrolases	Guanine deaminase			Native
ACJ52052.1	100%	GDSL-like Lipase/Acylhydrolase family	Lipase			Native
ACJ53073.1	100%	MFS multidrug transporter	MFS transporter	Tet38 tetracycline-resistance protein	2.A.1.3.22	Native
ACJ51151.1	100%	-	-			Foreign
ACJ53179.1	100%	RecG, C-terminal domain superfamily	Transcriptional regulator, partial			Native
ACJ52471.1	100%	-	-			Native
ACJ51149.1	100%	RelB antitoxin/Antitoxin DinJ	Type II toxin-antitoxin system family			Foreign
ACJ51552.1	100%	-	-			Native
ACJ53224.1	100%	-	5'-nucleotidase C-terminal domain			Native
ACJ51550.1	100%	Response regulatory domain profile.	Response regulator transcription factor			Native

ACJ51553.1	100%	nucleoside triphosphate hydrolases	ATP-binding domain-containing protein			Foreign
ACJ52096.1	100%	Lantibiotic protection ABC transporter permease	Lantibiotic immunity ABC transporter permease	3-component subtilin immunity exporter	3.A.1.124.2	Foreign
ACJ52097.1	100%	ABC-2 family transporter protein	Lantibiotic immunity permease	CprABC antimicrobial peptide resistance ABC exporter	3.A.1.124.6	Foreign
ACJ51554.1	100%	-	-			Native
ACJ51425.1	90%	Antitoxin	-			Foreign
ACJ51673.1	90%	ABC superfamily metabolite uptake	ABC transporter permease	Putative macrolide-specific efflux system, MacAB	3.A.1.122.16	Native
ACJ53183.1	90%	Metallophosphoesterase, calcineurin family	Metallophosphoesterase			Native
ACJ53406.1	90%	Sialidase	Exo-alpha-sialidase			Native
ACJ52932.1	90%	MFS_MefA_like	MFS transporter	The tetracycline resistance determinant, TetV	2.A.1.21.3	Native
ACJ52100.1	90%	HAMP domain-containing histidine kinase	-	-		Native
ACJ52095.1	90%	lantibiotic, protection ABC transporter ATP binding protein	-	-		Foreign
ACJ51226.1	90%	Protein/nucleic acid deglycase dj-1-related	DJ-1/PfpI family protein			Native
ACJ53416.1	90%	Beta-lactamase superfamily domain	MBL fold metallo-hydrolase			Native
ACJ53415.1	90%	PBP2_UgpB	ABC transporter substrate-binding protein	Involved in maltose and maltodextrin uptake	CEF11988.1	Native
ACJ51154.1	90%	ABC transporter integral membrane type-1	Phosphonate ABC transporter, permease	Putative phosphonate/phosphite/phosphate porter	3.A.1.9.2	Native
ACJ51426.1	90%	type II toxin-antitoxin system	BrnT family toxin			Native
ACJ51156.1	90%	ABC transporter-type domain profile	Phosphonate ABC transporter ATP-binding	Putative phosphonate/phosphite/phosphate porter	3.A.1.9.2	Native
ACJ53417.1	90%	Transport system inner membrane component	Carbohydrate ABC transporter permease	ABC transporters for maltose/maltotriose and trehalose	3.A.1.1.23	Native
ACJ51155.1	90%	phosphonate ABC transporter, permease	ABC transporter, permease protein	Putative phosphonate/phosphite/phosphate porter	3.A.1.9.2	Native
ACJ51158.1	90%	SIS_RpiR	MurR/RpiR family transcriptional regulator			Native
ACJ51157.1	90%	Periplasmatic phosphonate-binding protein	ABC transporter substrate-binding protein	Putative phosphonate/phosphite/phosphate porter, PhnDCE	3.A.1.9.2	Native
ACJ51374.1	90%	MFS_MdtG_SLC18_like	MFS transporter	Copper Uptake Porter	2.A.1.81.-	Native
ACJ51159.1	90%	HAD-like superfamily	HAD family hydrolase			Native
ACJ51153.1	90%	5'-Nucleotidase/apyrase	Metallophosphoesterase			Native
ACJ53414.1	90%	Haloacid dehalogenase-like hydrolase	HAD family hydrolase			Native
ACJ53419.1	90%	ABC transporter-type domain profile.	ABC transporter ATP-binding protein	Involved in the uptake of pectin oligosaccharides	3.A.1.1.34	Native
ACJ53418.1	90%	ABC transporter integral membrane type-1	Sugar ABC transporter permease	The fructooligosaccharide porter	3.A.1.1.20	Native
ACJ51575.1	90%	Glycosidases	Family 20 glycosylhydrolase			Native
ACJ51984.1	90%	MetI-like	Sugar ABC transporter permease	The xylobiose porter; BxlEFG(K)	3.A.1.1.21	Native

ACJ51567.1	90%	-	Tyrosine-type recombinase/integrase			Foreign
ACJ51985.1	90%	Maltose transport system permease	ABC transporter permease subunit	N-Acetylglucosamine/N,N'-diacetyl chitobiose porter	3.A.1.1.18	Native
ACJ53244.1	90%	Duplicated hybrid motif	PTS glucose transporter subunit IIA			Native
ACJ51983.1	90%	Carbohydrate substrate-binding protein	Carbohydrate ABC transporter	xylobiose porter	3.A.1.1.21	Native

B. longum* subsp. *suis

Core Gene	Prevalence across the subspecies	Function		Transporter Classification Database		HGT events
		Interpro Database	Refseq Database	Function	Family	
AIF90321.1	100%	ABC transporter, atp-binding protein	ABC transporter ATP-binding protein	The Macrolide Exporter (MacB) Family	3.A.1.122.-	Native
AIF90322.1	100%	ABC transporter permease	MacB-like periplasmic core domain	Exports macrolide antibiotics	3.A.1.122.18	Native
SDO34397.1	70%	Heavy metal transporter	ABC transporter ATP-binding protein			Native
SDO30658.1	70%	ABC-2 type transporter	FHA domain-containing protein	ABC exporter involved in bacterial competitiveness	3.A.1.105.19	Native
AIF89665.1	70%	vWA-like	VWA domain-containing protein			Native

Functional assessment of *B. longum* subspecies-specific core genomes

In order to gain further insight into the physiological characteristics of each *B. longum* subspecies, we investigated the *Bll*-CG, *Bls*-CG, and *Bli*-CG from a functional perspective by similarity searches in the NCBI RefSeq nr database [61] and protein domain prediction by InterProScan [62].

Of the 24 core genes unique to *B. longum* subsp. *longum,* eight could not be functionally annotated due to the absence of homologs with known function in the RefSeq nr database and known protein domains. In contrast, four were predicted to encode carbohydrate-utilization enzymes (Figure 2a).

Specifically, genes encoding pullulanase type I belonging to glycoside hydrolase family 13 (GH13), alpha-amylase (GH57), and a member of the amidohydrolase family proteins were found to be present in 97 % of the analyzed genomes. Moreover, a gene whose protein product resembles members of glycosyl hydrolase family 31, representing enzymes such as alpha-glucosidase, glucoamylase, alpha-xylosidase, and sucrase-isomaltase, was detected among 86 % of the analyzed *B. longum* subsp.

longum genomes (Table 1). Interestingly, the abovementioned enzymes are typically involved in the utilization of plant-related carbohydrates, which, being undigested by the host, are thus available as a carbon source by the microbiota resident in the colon. This finding corresponds with *B. longum* subsp. *longum* being commonly present in fecal samples of human adults, whose diet includes such plant carbohydrates [14].

The functional dissection of genes attributed to *Bls*-CG revealed the ubiquitous presence of two genes encoding ATP-binding cassette (ABC) transporters that are implicated in macrolide resistance. At the same time, among 90 % of the strains, an additional ABC transporter was found to be involved in the detoxification of heavy metals (Table 1, Figure 2a), allowing us to infer that these transporters play a critical role in niche adaptation of this subspecies. Notably, macrolides have been reported to be among the most frequently used classes of antimicrobials in pig breeding [63]. Therefore, they may have facilitated the development of antimicrobial resistance in bacteria present in the porcine gut microbiota [64].

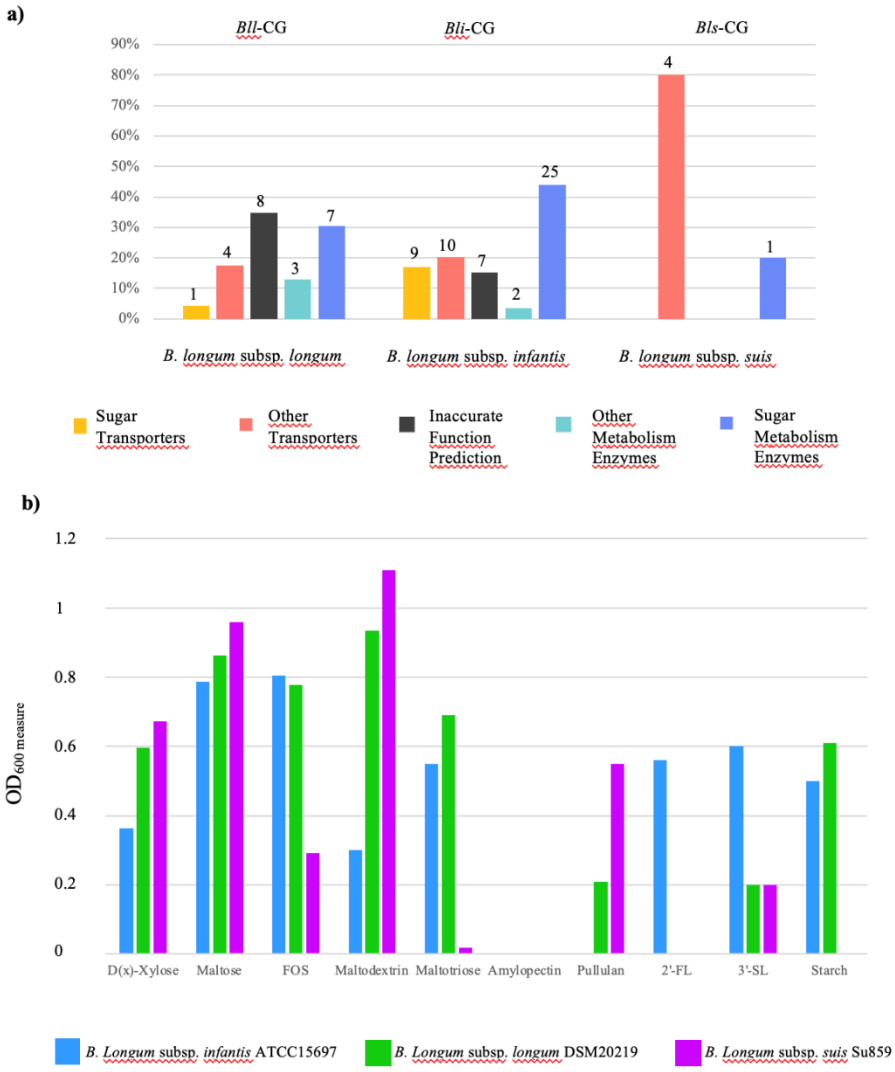


Figure 2. Functional annotation of core genes and growth performance of *B. longum* subspecies. Panel **(a)** shows the distribution in functional categories of the core genes identified in each *B. longum* subspecies. The number of genes assigned to each category is reported at the top of the columns. Panel **(b)** displays the growth performance of each *B. longum* type strain on different carbohydrates expressed through measurement of OD_{600nm}.

Focusing on the *B. longum* subsp. *infantis*-specific core genetic repertoire, the most noticeable difference with respect to *B. longum* subsp. *longum* is the presence of genes involved in transporting a broad range of carbohydrates (Figure 2a). Specifically, all assessed *B. longum* subsp. *infantis* genomes encompass genes that

are predicted to encode a glucose transporter and a glucose/galactose-Na^+ symporter. Interestingly, since glucose and galactose are the building blocks constituting lactose through glycosidic linkage, this finding suggests the presence of an extracellular (bifido)bacterial β-galactosidase (GH42), as also reported in previous genomic surveys [65], and improved specialization toward the uptake of released simple sugars. Moreover, we also identified transporters related to Tetracycline resistance, including the tetracycline resistance determinant Tet(V) and *Tet38* gene. Tetracyclines are one of the most widely used groups of antibiotics worldwide, and resistance to this class of antibiotics is widespread even among bacteria that colonize the infant gut [66, 67]. Therefore, it may represent a trait that increases competitiveness of *B longum* subsp. *infantis* in the gastrointestinal tract.

Progressively extending the analysis of *Bli*-CG by decreasing the level of gene sharing amongst members of this subspecies, i.e., prevalence, we identified genes that were involved in the uptake of pectic oligosaccharides (POSs), fructooligosaccharides (FOSs), maltose/maltotriose, and xylobiose with a prevalence of 90 % (Table 1).

Intriguingly, these observations corroborated the well-known bifidogenic properties exhibited by POSs and FOSs, routinely used in commercial prebiotics due to their beneficial impact on the gut microbiota [68, 69]. Furthermore, within 90 % of *B. longum* subsp. *infantis* genomes, we identified an exo-alpha sialidase (GH33) and a member of glycosyl hydrolase family 20 (GH20) (Table 1), which represent enzymes known to be implicated in the metabolism of HMOs. These latter glycans are not processed by human digestive enzymes, thus reaching the colon intact where they are metabolized by certain members of the resident microbial community, such as *B. bifidum*, *B. breve* as well as *B. longum*, which encode gene clusters specifically dedicated to HMO metabolism [13, 70].

In particular, sialidases (GH33) catalyze the removal of terminal sialic acid residues, thus playing a critical role in the degradation of sialylated HMOs, such as 3′- and 6′-sialyllactose [71]. In line with previous publications, these findings show that degradation of sialylated HMOs is an ability that seems to be distinctive for *B. longum* subsp. *infantis*, thus being a characterizing genotypic and phenotypic feature of this subspecies [15]. In contrast, the above described GH20 family comprises enzymes with β-hexosaminidase and lacto-N-biosidase activities, which act on substrates that form part of the HMO backbone, thereby releasing N-acetylglucosamine and lactose molecules, respectively [15, 72].

Previous investigations of the bifidobacterial glycobiome have highlighted that most *B. longum* subsp. *longum* strains (75-100% of the strains) are predicted to encode the β-hexosaminidase (GH20), the lacto-N-biose phosphorylase (GH112), as well as an extracellular lacto-N-biosidase (GH136) [14, 73]. Intriguingly the finding of additional genes belonging to GH20 family encoded only by members of the *B. longum* subsp. *infantis* suggests that this subspecies has been subject to specific evolutionary selection. Remarkably, the latter seems to have driven *B. longum* subsp. *infantis* towards the acquisition of HMO-metabolizing genes, in addition to those shared with other members of the *B. longum* species.

Overall, the observed uneven distribution of the carbohydrate-active enzyme arsenal may reflect the distinct colonization strategy adopted by each *B. longum* subspecies, indicating that *B. longum* subsp. *longum* is more adapted to a (human) adult diet, as also supported by previous findings [15]. In contrast, members of the *B. longum* subsp. *infantis* subspecies may have evolved from a plant-derived glycan utilization gene-makeup towards a genomic repertoire that aims to achieve efficient colonization of the suckling mammalian gut.

To validate these *in silico* results, which indicate a more dedicated commitment of *B. longum* subsp. *longum* toward the breakdown of plant-related carbohydrate when

compared to *B. longum* subsp. *infantis*, growth of the type strains of each *B. longum* subspecies, namely *B. longum* subsp. *longum* DSM20219, *B. longum* subsp. *suis* DSM20097, and *B. longum* subsp. *infantis* ATCC15697, was evaluated on ten different carbohydrates. In detail, for growth profiling experiments, we used a carbohydrate-free basic MRS medium which was supplemented with either amylopectin, pullulan, starch, maltotriose, maltodextrin, xylose, 2'-FL, 3'-SL, FOS, or maltose as the sole carbon source (Table S4, Figure 2b).

Based on our analyses, *B. longum* subsp. *suis* was the only subspecies able to grow on pullulan-based medium (final OD above 0.5). Appreciable growth was also observed on xylose, maltose, and maltodextrin (final OD ranging from 0.67 to 1.11). Conversely, both *B. longum* subsp. *longum* and *B. longum* subsp. *infantis* was shown to be able to grow on starch and starch-like glycans (final OD above 0.5), with the exception of amylopectin and pullulan, for which no appreciable growth was noticed (Table S4, Figure 2b). Nevertheless, as is displayed in Figure 2b, *B. longum* subsp. *infantis* was shown to exhibit a reduced level of metabolic abilities on various assessed plant-related glycans when compared to those elicited by *B. longum* subsp. *longum.* Furthermore, *B. longum* subsp. *longum* was shown to possess the most elaborate plant-related carbohydrate degrading activities among the *B. longum* species, being consistent with the above described *in silico* reports (Table S4, Figure 2b) [74].

Furthermore, *B. longum* subsp. *infantis* appears to be the only subspecies type strain capable of metabolizing 2'-FL and 3'-SL (Table S4, Figure 2b). Consistently, the pronounced ability of *B. longum* subsp. *infantis* to metabolize a wide range of HMO compounds has been extensively reported [75-77]. However, carbohydrate metabolism data available in the literature have also highlighted specific HMO-utilizing abilities for certain members of the *B. longum* subsp. *longum*. While all strains can efficiently metabolize lacto-N-tetraose (LNT) and lacto-N-biose (LNB),

only certain strains have shown growth capabilities on fucosylated HMOs and Lacto-N-neotetraose (LNnt) [76, 78]. In fact, growth profiles of the latter subspecies resemble that of *Bifidobacterium adolescentis* [74], which represents a gut-resident bifidobacterial taxon typical of the post-weaning period [79].

Overall, the findings related to the *in vitro* growth experiments corroborate our *in silico* data and may be a reflection of the ecological niche in which each *B. longum* subspecies dominates. Our data therefore suggest that *B. longum* subsp. *longum* plays an ecological role in the metabolism of dietary, plant-derived carbohydrates during weaning and post-weaning phases when infants are gradually introduced to a solid diet containing such complex carbohydrates [80-82]). Accordingly, the identified fermentation capabilities may provide an explanation as to how *B. longum* subsp. *longum* is able colonize both the infant and adult gut. In contrast, *B. longum* subsp. *infantis* is more adapted to colonization of the pre-weaning gut environment due to its particular HMO degradation abilities [15, 83].

Mobilome prediction in *B. longum* genomes

Horizontal Gene Transfer is the process by which genetic material is exchanged between and within microbial taxa/taxon [84, 85]. This phenomenon of acquisition of new genomic properties is crucial for adaptation to new ecological niches [86], while it generates genetic diversity across bacterial taxa [87]. To a large degree, among (bifido)bacteria, HGT is assumed to occur through mobile genetic elements, such as plasmids, transposons or bacteriophages, with the latter considered one of the main vectors for gene transfer [88, 89]. To explore the possibility that HGT events are responsible for the substantial intra-specific genomic diversity observed between *B. longum* subspecies, the genomes of the representative 42 strains previously selected for phylogenetic analyses (Figure 1) were screened using the software Colombo [41].

Following bioinformatic inspection of the *B. longum* subsp. *longum* and *B. longum* subsp. *suis* genomes, an average of 431 and 407 putative HGT genes, corresponding to an average of 22.5 % and 20.9 % of the total number of CDS, respectively, were identified (Figure 3a, Table S5). In contrast, an average of 640 CDS, corresponding to 29.5 % of the total number of predicted CDS, were identified as being potentially acquired by HGT in *B. longum* subsp. *infantis* (Figure 3a, Table S5). To get an idea of the extent to which HGT events have contributed to shaping the genome architecture of *B. longum* subspecies, these values were compared to those obtained from 85 type strains belonging to different bifidobacterial species. Overall, the latter genomes showed an average of 12.8 % putative HGT-acquired genes, which was significantly lower than those identified in *B. longum* subspecies (ANOVA p-value < 0.01) (Figure 3a, Table S6), as was previously reported [90].

Furthermore, it is particularly noteworthy that *B. longum* subsp. *infantis* elicits the highest HGT gene numbers among the assessed bifidobacterial (sub)species, highlighting that this subspecies appears to be more suitable or to have been subject to higher selective pressure to acquire alien DNA when compared to not only other bifidobacterial species, but also compared to other *B. longum* subspecies (ANOVA p-value < 0.01). Accordingly, these results provide an explanation for the higher average genome size of the *B. longum* subsp. *infantis* chromosomes.

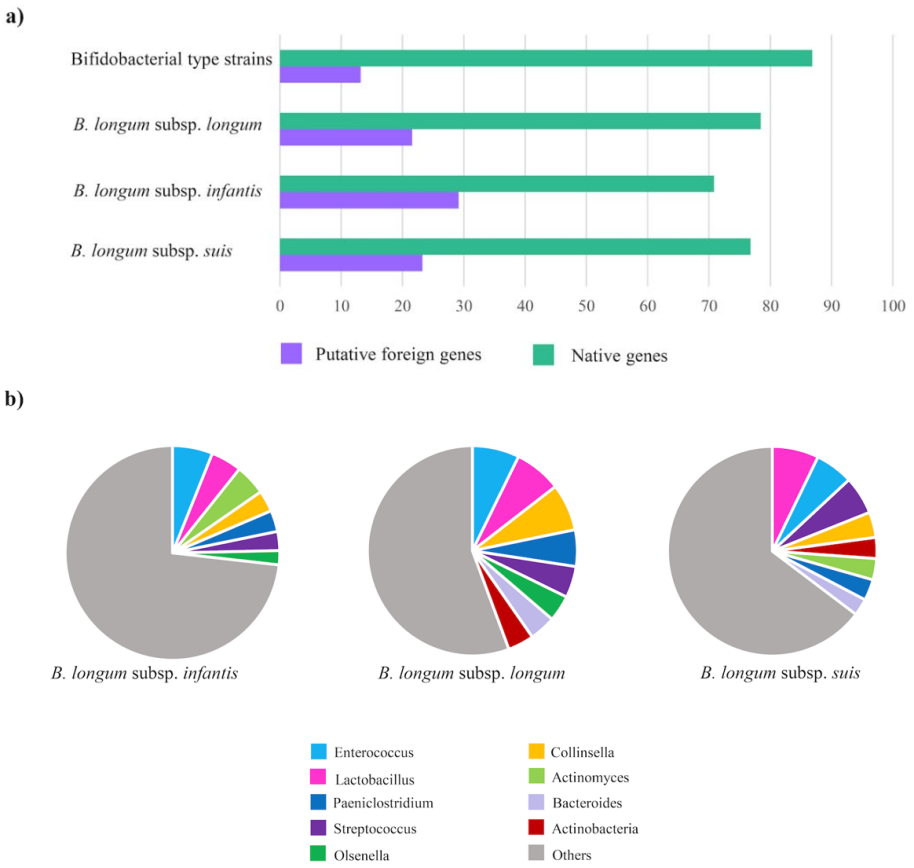


Figure 3. Prediction of *B. longum* subspecies HGT events. Panel **(a)** shows the average percentages of the predicted foreign genes in each *B. longum* subspecies, compared with those obtained from 85 type strains belonging to different bifidobacterial species. Panel **(b)** displays the predominant non-bifidobacterial donor genera of the putative alien genes found in each *B. longum* subspecies.

Subsequently, the genes predicted to have been horizontally acquired by each of the *B. longum* subspecies were subjected to similarity searches in the NCBI refseq nr database in order to obtain an overview of the potential donor taxa. In particular, 124 (35 %) of the identified foreign genes of *B. longum* subsp. *longum*, 153 (30 %) of those of *B. longum* subsp. *suis*, and 280 (44 %) of the alien genes detected in *B. longum* subsp. *infantis* returned significant database hits in terms of similarity. Interestingly, of these identified HGT genes, 118 (95 %) of *B. longum* subsp. *longum*, 124 (81 %) of *B. longum* subsp. *suis,* and 191 (68 %) of *B. longum* subsp. *infantis*,

corresponding respectively to 27 %, 30 %, and 30 % of the total HGT-acquired genes, appear to be derived from other bifidobacterial species, most frequently by *B. bifidum*, *B. breve,* and *Bifidobacterium adolescentis* (Tables S7-S9). These latter species are also commonly found in the gastrointestinal tract of infants, thus representing a common niche that would facilitate horizontal transfer events. Furthermore, following the exclusion of hits corresponding to genera belonging to the Bifidobacteriaceae family, the analysis revealed a preferential origin of alien DNA from *Enterococcus*, *Lactobacillus*, *Streptococcus*, *Collinsella*, *Bacteroides, Actinomyces* as well as *Paeniclostridium* (Tables S7-S9).

In particular, *Enterococcus* (7.2 %), *Lactobacillus* (7.2 %), and *Collinsella* (7.2 %) were identified as major donors of the *B. longum* subsp. *longum* horizontal genes (Figure 3b, Table S8), while *Lactobacillus* (7.1 %)*, Enterococcus* (5.8 %), and *Streptococcus* (5.8 %) were recognized as the prominent donors of the *B. longum* subsp. *suis* foreign genes (Figure 3b, Table S9). In a similar fashion, the *B. longum* subsp. *infantis* genes putatively acquired by HGT were predicted to be originated mainly from *Enterococcus* (6.1 %), *Lactobacillus* (4.6 %) and *Actinomyces* (4.6 %) (Figure 3b, Table S7). Interestingly, these donor genera, including the bifidobacterial ones, are known to share the human (infant) gut environment with members of *B. longum* species [7, 91], thus providing the opportunity for genetic transfer events, which can act as the driver of niche adaptation in members of the *B. longum* species [92].

To further investigate how HGT can contribute to differentially shape the *B. longum* subspecies, we assessed to what extent potential HGT events affect the specific core genome of each *B. longum* subspecies (Table 1). Notably, we found two alien core genes in the *Bll*-CG, and 10 putative alien genes in the *Bli*-CG, corresponding respectively to 8.3 % and 18.8 % of their own total number of core genes. Instead,

no horizontal core genes were found among the five constituting the *Bls*-CG (Table 1).

As expected, HGT seems to contribute only marginally to the core genome of the three *B. longum* subspecies. This observation is consistent with the notion that core genes are the most ancient genes, whose acquisition shaped the ancestors of each *B. longum* subspecies [93, 94]. Nevertheless, *B. longum* subsp. *infantis* was predicted to possess a higher number of foreign core genes compared to the other *B. longum* subspecies. Furthermore, based on RefSeq database annotation, horizontally acquired core genes in *Bli*-CG encompassed five genes putatively involved in the production of antimicrobial peptides, such as bacteriocins, and genes related to a toxin/antitoxin system (Table 1). Notably, bacteriocins are commonly produced by lactic acid bacteria, including *Lactobacillus*, *Streptococcus*, and members of the *Enterococcus* genus [95, 96] that were consistently found amongst the major donor genera of foreign *B. longum* genes.

Survey of the genetic features supporting HGT events

Mobile genetic elements, such as transposable elements and prophage-like elements, can promote DNA acquisition and facilitate the genetic material transmission between different bacterial taxa [87]. Conversely, Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and restriction-modification (R-M) systems, which both represent microbial defense mechanisms against invasion of alien genetic material, are responsible for the degradation of nonself-DNA thereby preventing HGT events [97]. In order to investigate the genetic features of *B. longum* subspecies involved in the acquisition of foreign DNA, the representative 42 genomes were screened for R-M and CRISPR-Cas systems (Figure 4a-b, Table S10).

Figure 4. R-M and CRISPR-Cas systems in *B. longum* subspecies. Panel **(a)** shows the number of genomic R-M systems found in each of the 58 representative *B. longum* strains. Panel **(b)** depicts the presence (orange) or absence (black) of CRISPR-Cas systems in each of the 58 representative *B. longum* strains.

Overall, these analyses revealed that *B. longum* genomes mainly harbor type II and type I R-M systems, with a higher average number of R-M enzymes found in the subspecies *B. longum* subsp. *longum*. In detail, this latter subspecies exhibited an average of 2.2 ± 0.8 R-M genes (Figure 4a, Table S10), while assessments of *B. longum* subsp. *suis* and *B. longum* subsp. *infantis* revealed the presence of 1.7 ± 1.1

and 0.45 ± 0.5 R-M genes, respectively (Figure 4a, Table S9). Interestingly, these results negatively correlate with the number of alien genes as mentioned above for each *B. longum* subspecies (t-test *p*-value < 0.05), corroborating the hypothesis that R-M systems counteract HGT events.

As mentioned above, CRISPR-Cas systems represent another bacterial defence mechanism against invading alien DNA [98]. Based on our screening, out of the 21 representative *B. longum* subsp. *longum* strains considered, 11 were shown to contain at least one complete CRISPR-Cas locus in their genome (prevalence of 52.4 %) (Figure 4b, Table S10). Besides, complete CRISPR-Cas systems were detected in 3 out of the 10 *B. longum* subsp. *suis* genomes, corresponding to a prevalence of 30 %, as well as within 3 out of the 11 *B. longum* subsp. *infantis* chromosomes, corresponding to a prevalence of 27 % (Figure 4b, Table S10). Furthermore, the screening highlighted the occurrence of type I (subtypes I-C, I-E, and I-U) and type II systems (subtypes II-C), characterized by the presence of *cas3* and *cas9* genes, respectively [99].

Specifically, a type II CRISPR-Cas system was detected only among *B. longum* subsp. *longum* strains, while such a system seems to be absent in *B. longum* subsp. *suis* and *B. longum* subsp. *infantis*. Overall, profiling of defense mechanisms highlighted that *B. longum* subsp. *longum* genomes seem to be equipped with a more efficient defense against foreign DNA invasion compared to those of both *B. longum* subsp. *suis* and *B. longum* subsp. *infantis* [100].

To obtain an overview of the *B. longum* genetic elements that may be implicated in HGT events, we screened for the presence of prophage-like and IS elements as well as plasmid sequences (Table S10). This allowed the identification of 21 (average of 1 per genome) and 8 (average of 0.8 per genome) prophage-like sequences in the inspected *B. longum* subsp. *longum* and *B. longum* subsp. *suis* genomes, respectively. In contrast, 22 (bifido)prophages, corresponding to an average of 2 integrated phages

per genome, were observed in the chromosomes of *B. longum* subsp. *infantis* (Table S10).

The genomic structural features of these retrieved bifidoprophages suggest that they represent members of the *Siphoviridae* family, consisting of lysogeny, DNA replication, DNA packaging, head and tail synthesis, and host lysis modules (Table S11). Furthermore, on average, 26.3 ± 11.8 and 23.4 ± 8.8 transposase genes per genome were found by inspecting *B. longum* subsp. *longum* and *B. longum* subsp. *suis* genomes, respectively, while *B. longum* subsp. *infantis* harbors 24.2 ± 19.9 transposase genes per genome. In contrast, *in silico* prediction did not reveal any plasmid sequences among the inspected *B. longum* genomes.

Altogether, these results, coupled with data obtained from the analysis of HGT occurrences, suggests that *B. longum* subsp. *infantis* seems to be more prone to acquire alien genes than the other two *B. longum* subspecies, highlighting how HGT events may have played a prominent role in shaping the genome of this taxon, ultimately providing it with specific ecological niche adaptations.

Conclusions

We investigated the genome diversity of *B. longum* species and its subspecies *B. longum* subsp. *longum, B. longum* subsp. *infantis* and *B. longum* subsp. *suis* through comparative genomic analyses and phylogenomic reconstruction of 261 publicly available and high-quality genomes, along with 11 novel strains sequenced as part of this study. These analyses revealed that members of *B. longum* subsp. *infantis* appear to contain a more extensive genetic repertoire than the other *B. longum* strains, highlighting how the former was shaped over the course of evolution through the acquisition of new genetic features. Notably, the functional analyses of the core genome unveiled that members of *B. longum* subsp. *infantis* possess unique

carbohydrate utilization capabilities toward host glycans, particularly those for HMO degradation.

When we investigated to what extent HGT events had been responsible for shaping *B. longum* subsp. *infantis* genomes, we revealed the increased frequency by which *B. longum* subsp. *infantis* had acquired alien DNA when compared to the other *B. longum* subspecies and to the type strains of other known bifidobacterial species. Notably, such higher genome plasticity, supported by specific genetic features such as lower number of restriction/modification and CRISPR-Cas systems coupled with a higher occurrence of prophage-like elements, appears to be the main reason that allowed *B. longum* subsp. *infantis* to adapt to early life mammalian gut colonization. Furthermore, prediction of putative donor taxa of alien DNA revealed a preferential origin from other bifidobacterial and non-bifidobacterial species inhabiting the gut environment, suggesting that the extensive milk-related carbohydrate utilization capabilities that characterize the *B. longum* subsp. *infantis* subspecies seems to have been obtained through extensive gene harvesting from co-colonizing bacterial taxa. Though our findings provide insights into how the three *B. longum* subspecies developed through differential gene acquisition and subsequent niche occupation, it should be kept in mind that our conclusions are predominantly based on bioinformatic analyses. Our future efforts will therefore aim to further support these *in silico* data with experimental evidence.

Nevertheless, certain limitations of this study should be kept in mind. In particular, the fact that the number of publicly available sequenced chromosomes belonging to *B. longum* subsp. *infantis* and *B. longum* subsp. *suis* is significantly lower compared with that of *B. longum* subsp. *longum* subspecies. This may imply that the genetic variation within the first two subspecies may not have been completely disentangled, as also demonstrated by the identification of an open pan-genome characterizing *B. longum* subsp. *infantis* as well as *B. longum* subsp. *suis* (see Supplementary Text).

In addition, our study very much focused on the *in silico* assessment of the genetic traits distinguishing each *B. longum* subspecies, highlighting the need for experimental validation of our presented bioinformatics data.

MATERIALS AND METHODS

Ethical statement

Animal research was performed in compliance with the rules, regulations and recommendations of the Ethical Committee of the University of Parma. The corresponding protocols were approved by the 'Comitato di Etica Università degli Studi di Parma', Italy. All animal procedures were carried out in accordance with national guidelines (Decreto legislativo 26/2014).

Furthermore, the human study protocol was approved by the Ethics Committee of the 'Azienda Unità Sanitaria Locale di Reggio Emilia - IRCCS' in Reggio Emilia, Italy, as well as by the Ethics Committee of the University of Parma, Italy, and informed written consent was obtained from all participants or their legal guardians.

B. longum genome sequences

At the time of writing (November 2020), 363 publicly available *B. longum* genomes (complete and draft genome sequences) were retrieved from the National Center for Biotechnology Information (NCBI) public database and then subjected to genome quality-based selection. In detail, genome sequences showing a genome size less than 2.20 Mb or/and with a number of predicted CDSs less than 1600 as well as those exhibiting low sequencing quality (genome coverage lower than 30-fold or containing more than 100 contigs) were manually identified and discarded. Furthermore, duplicated bacterial genomes (ANI value > 99.99 %) were removed, resulting in a final collection of 261 high-quality *B. longum* genomes encompassing

243, seven, and 11 chromosomes belonging to *B. longum* subsp. *longum*, *B. longum* subsp. *suis*, and *B. longum* subsp. *infantis* subspecies, respectively. Furthermore, we decoded the chromosomes of 11 newly isolated *B. longum* strains that were also included in this study (Table S1). Notably, these latter isolates were obtained from human, bovine, and canine fecal samples within the context of a bifidobacterial strain isolation project aimed at exploring the genetic variability of the *Bifidobacterium* genus.

Identification of novel *B. longum* strains and chromosomal DNA extraction

Based on a previous cultivation effort aimed at isolating *Bifidobacterium pseudolongum* strains from fecal samples of various mammalian species [29], bifidobacterial strains that did not belong to the above mentioned bifidobacterial species were further subjected to species-specific PCR-based characterization in order to identify novel *B. longum* strains. Briefly, bifidobacterial strains were incubated in an anaerobic atmosphere (2.99 % H_2, 17.01 % CO_2 and 80 % N_2) in a chamber (Concept 400, Ruskinn) in de Man-Rogosa-Sharpe (MRS) (Sharlau Chemie) supplemented with 0.05 % (wt/vol) L-cysteine hydrochloride and incubated at 37°C for 16 hours. Subsequently, cells were harvested by centrifugation at 3,500 x *g* for 8 min, and the obtained cell pellet was used for DNA extraction using the GenEluteTM Bacterial Genomic DNA kit (Merck, Germany), following the manufacturer's instructions. The extracted DNA was then subjected to a *B. longum* species-specific identification protocol through a PCR-based methodology using primers Blong1 5'-TCCCAGTTGATCGCATGGTC-3' and Blong2 5'-GGGAAGCCGTATCTCTACGA-3', which are based on the 16S rRNA gene sequences of this taxon [30]. PCR amplification was carried out according to the following protocol: one cycle of 94 °C for 5 min, followed by 30 cycles of 94 °C for 30s s, 54 °C for 30 s and 72 °C for 50 s, and a final cycle of 72 °C for 5 min.

Furthermore, the DNA of strains identified as *B. longum* ssp. were further subjected to a genotyping PCR using primers ERIC1 5'-ATGTAAGCTCCTGGGGATTCAC-3' and ERIC2 5'-AAGTAAGTGACTGGGGTGAGCG-3' in order to sequence the genome of only one representative per genotype [31]. PCR amplification was performed according to a previous protocol: one cycle at 94 °C for 3 min, followed by 35 cycles of 94 °C for 30 s, 48 °C for 30 s and 72 °C for 4 min, and a final cycle at 72 °C for 10 min [31].

B. longum genome sequencing and assemblies

Chromosomal DNA of the 11 newly identified *B. longum* strains was sequenced by GenProbio Srl (http://genprobio.com) using a MiSeq platform (Illumina, San Diego, CA, USA) according to the supplier's protocol employing the Nextera XT DNA Library Prep Kit (Illumina), resulting in fragments of about 500-900 bp. The library samples obtained were then pooled into a Flow Cell V3 600 cycle (Illumina) in order to retrieve paired-end reads of 250 bp resulting from sequencing of fragment ends. Fastq files of paired-end reads generated from each genome sequencing effort were used as input for the genome assembly through the MEGAnnotator pipeline (https://github.com/GabrieleAndrea/MEGAnnotator) [32]. The SPAdes v3.14.0 program included in the MEGAnnotator platform was used for *de novo* assembly of each bifidobacterial genome sequence with the pipeline option "--careful" and a list of k-mer sizes 21,33,55,77,99,127 as suggested in the SPAdes' manual [33]. MEGAnnotator then employed contigs greater than 1000 bp to predict protein-encoding open reading frames (ORFs) using Prodigal v2.0 [34]. Predicted ORFs were then functionally annotated using RAPSearch2 (reduced alphabet based protein similarity search) (cutoff e-value of 1×10^{-5} and minimum alignment length 20) employing the NCBI reference sequences (RefSeq) database [35] together with

hidden Markov model profile (HMM) searches (http://hmmer.org/) performed against the manually curated Pfam-A database (cutoff e-value of 1×10^{-10}).

Pan-genome analyses of *B. longum* genomes

All 272 genome sequences of *B. longum* were employed for a pan-genome analysis using the Pangenome Analysis Pipeline (PGAP) v1.1 [36] (http://pgap.sf.net). Predicted CDSs of each *B. longum* genome were classified into functional gene clusters through the gene family (GF) method, consisting of pairwise protein-similarity search employing BLAST software v2.2.28+ (cutoff e-value of 1×10^{-10} and exhibiting at least 50 % identity across at least 80 % of both protein sequences). Following this, using MCL (graph-theory-based Markov clustering algorithm) [37], the data obtained were used to assign proteins to so-called Clusters of Orthologous Groups (COGs). A pan-genome profile was then built using an optimized algorithm as part of the PGAP software v1.1, based on a presence/absence matrix encompassing all COGs identified in the analyzed genomes (Linux command line "./PGAP.pl --strains [input_strain_list] --input input_path/ --output output_path/ --thread 20 --identity 0.5 --coverage 0.8 --cluster --method GF --evolution --pangenome"). Subsequently, the core genome of *B. longum* species was obtained by selecting protein families which are shared between all genomes, while truly unique genes (TUGs) encoded by a single genome were identified based on those protein families that are present in one *B. longum* genome yet absent in all other *B. longum* genomes. Separate pan- and core- genome analyses were performed on each *B. longum* subspecies as described above, involving genomes of 251 *B. longum* subsp. *longum*, 11 *B. longum* subsp. *infantis* and ten *B. longum* subsp. *suis* genomes.

Phylogenomic comparison between *B. longum* strains

In order to assess the genetic relatedness among the 272 members of *B. longum* species, the COGs constituting the core genome of each *B. longum* strain were concatenated, and they were then aligned using MAFFT v7.222 [38] through the Linux command line "mafft --thread 20 --retree 2 --clustalout --reorder [input_sequences.fasta] > output.aln". The resulting phylogenomic tree was constructed using the neighbor-joining method in ClustalW v2.1 [39] through the Linux command line "clustalw -bootstrap=100 -seed=100 -bootlabels=NODE -outputtree=phylip -infile=file.aln". Then, utilizing the graphical viewer of phylogenetic trees FigTree v1.4 (http://tree.bio.ed.ac.uk/software/figtree/), the core genome-based visual tree was developed. Furthermore, a value for the average nucleotide identity (ANI) was calculated for each genome pair using the fastANI software v1.3 [40] through the Linux command line "./fastANI --ql [genome_list_path] --rl [genome_list_path] -t 20 --matrix -o output.txt". Out of 272 obtained *B. longum* genomes, we selected 42 *B. longum* strains in order to perform downstream analyses (Figure 1). For this purpose, we included all ten genomes that clustered with the *B. longum* subsp. *suis* type strain DSM20097 (seven publicly available and three newly isolated), 11 of the non-redundant identified *B. longum* subsp. *infantis* chromosomes with suitable quality (see above), along with 21 representative of *B. longum* subsp. *longum*. Notably, these latter comprised the type strain DSM 20219, eight newly isolated, and an additional 12 publicly available genome sequences, selected to maximize the description of the intra-subspecies diversity from the branch of the tree encompassing the whole selection of *B. longum* subsp. *longum* (Figure S2).

Prediction of the mobilome of *B. longum*

The identification of the genes that may have been acquired by Horizontal Gene Transfer (HGT) events was achieved using the suite COLOMBO v3.8, with a sensitivity value of 0.7 (https://github.com/brinkmanlab/colombo/releases) [41]. Furthermore, the proteome of each *B. longum* strain was screened for the presence of Restriction-Modification (R-M) systems based on sequence similarity to genes classified in the REBASE database [42] (http://rebase.neb.com/rebase/rebhelp.html; BLAST cutoff e-value of 1×10^{-5}). The presence of transposable elements was performed through the IS Finder online tool with predefined parameters (https://isfinder.biotoul.fr/), while identification of clustered regularly interspaced short palindromic repeats (CRISPRs) was achieved through the web application CRISPRfinder (https://crispr.i2bc.paris-saclay.fr/Server/; default parameters were used) [43]. Prediction of prophage-like elements was conducted using a custom BLAST database (cutoff e-value of 1×10^{-5}) encompassing previously bifidoprophage-validated sequences obtained from bifidobacterial type strains previously described [44]. Then, genomic regions encompassing predicted phage-related genes were manually examined to identify complete prophage-like sequences. Assessment of complete or partial plasmid sequences was carried out employing a combination of the PlasmidFinder 2.1 web service (https://cge.cbs.dtu.dk/services/PlasmidFinder/ ; minimum identity = 50 % and minimum coverage = 80 %) [45] and ABRicate software (https://github.com/tseemann/abricate).

***B. longum* type strains carbohydrate growth assays**

In order to validate the *in silico* findings, we performed growth assays on selected carbon sources involving the type strains of each *B. longum* subspecies, i.e., *B. longum* subsp. *longum* DSM20219, *B. longum* subsp. *suis* DSM20097, and *B.*

longum subsp. *infantis* ATCC15697. Notably, *in silico* analyses performed in this study generated predictions with regards to (carbohydrate) metabolic abilities of the abovementioned strains and further discussed in the Results section. *B. longum* type strains were cultivated overnight on semisynthetic MRS medium supplemented with 0.05 % (w/vol) L-cysteine hydrochloride at 37°C under anaerobic conditions. Subsequently, cells were diluted in MRS without glucose in order to obtain an OD_{600nm}=1 and 15 µl of the diluted cells were inoculated in 135 µl of MRS without glucose supplemented with 1 % (wt/vol) of a particular sugar in a 96-well microtiter plate, and incubated in an anaerobic cabinet. Specifically, each carbohydrate was dissolved in MRS without glucose previously sterilized by autoclaving at 121 °C for 15 min. Subsequently, each obtained solution was filter sterilized using a 0.2 µm filter size prior to use. Cell growth was evaluated by monitoring the optical density at 600 nm with the use of a plate reader (Biotek, VT, USA). The plate was read in discontinuous mode, with absorbance readings performed at 3 min intervals for three times after 48 h of growth, and each reading was ahead of 30 s of shaking at medium speed. Cultures were grown in triplicates, and the resulting growth data were expressed as the average of these replicates. Carbohydrates tested in this study were purchased from Merck (Germany) and Carbosynth (Berkshire, United Kingdom), and include soluble starch from potato, amylopectin from maize, pullulan, maltotriose, maltodextrin, FOS, D-(+)-maltose, D-(+)-xylose, 2'-Fucosyllactose (2'-FL), 3'-Sialyllactose (3'-SL), and α-D-glucose.

Statistical analyses

All statistical analyses were performed with SPSS software v25 (www.ibm.com/software/it/analytics/spss/).

Conflicts of interest

The authors declare that there are no conflicts of interest.

Funding information

This work received no specific grant from any funding agency.

Acknowledgments

GA is supported by Fondazione Cariparma, Parma, Italy. We furthermore thank GenProbio Srl for the financial support of the Laboratory of Probiogenomics. This research benefited from the HPC (High-Performance Computing) facility of the University of Parma, Italy. D.v.S. is a member of APC Microbiome Ireland, which is supported by Science Foundation Ireland, through the Irish Government's National Development Plan (SFI/12/RC/2273-P1 and SFI/12/RC/2273-P2).

References

1. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A 1998;95(12):6578-6583.

2. Donovan SM. Introduction to the special focus issue on the impact of diet on gut microbiota composition and function and future opportunities for nutritional modulation of the gut microbiome to improve human health. Gut Microbes 2017;8(2):75-81.

3. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ et al. Metagenomic analysis of the human distal gut microbiome. Science 2006;312(5778):1355-1359.

4. Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI. Host-bacterial mutualism in the human intestine. Science 2005;307(5717):1915-1920.

5. Milani C, Duranti S, Bottacini F, Casey E, Turroni F et al. The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. Microbiol Mol Biol Rev 2017;81(4).

6. Turroni F, Milani C, Duranti S, Ferrario C, Lugli GA et al. Bifidobacteria and the infant gut: an example of co-evolution and natural selection. Cell Mol Life Sci 2018;75(1):103-118.

7. Mancabelli L, Tarracchini C, Milani C, Lugli GA, Fontana F et al. Multi-population cohort meta-analysis of human intestinal microbiota in early life reveals the existence of infant community state types (ICSTs). Comput Struct Biotechnol J 2020;18:2480-2493.

8. Benno Y, Sawada K, Mitsuoka T. The intestinal microflora of infants: composition of fecal flora in breast-fed and bottle-fed infants. Microbiol Immunol 1984;28(9):975-986.

9. Hidalgo-Cantabrana C, Delgado S, Ruiz L, Ruas-Madiedo P, Sanchez B et al. Bifidobacteria and Their Health-Promoting Effects. Microbiol Spectr 2017;5(3).

10. Ruiz L, Delgado S, Ruas-Madiedo P, Sanchez B, Margolles A. Bifidobacteria and Their Molecular Communication with the Immune System. Front Microbiol 2017;8:2345.

11. Turroni F, Bottacini F, Foroni E, Mulder I, Kim JH et al. Genome analysis of Bifidobacterium bifidum PRL2010 reveals metabolic pathways for host-derived glycan foraging. Proc Natl Acad Sci U S A 2010;107(45):19514-19519.

12. Smilowitz JT, Lebrilla CB, Mills DA, German JB, Freeman SL. Breast milk oligosaccharides: structure-function relationships in the neonate. Annu Rev Nutr 2014;34:143-169.

13. James K, Motherway MO, Bottacini F, van Sinderen D. Bifidobacterium breve UCC2003 metabolises the human milk oligosaccharides lacto-N-tetraose and lacto-N-neo-tetraose through overlapping, yet distinct pathways. Sci Rep 2016;6:38560.

14. Lugli GA, Duranti S, Milani C, Mancabelli L, Turroni F et al. Investigating bifidobacteria and human milk oligosaccharide composition of lactating mothers. FEMS Microbiol Ecol 2020;96(5).

15. Sela DA, Chapman J, Adeuya A, Kim JH, Chen F et al. The genome sequence of Bifidobacterium longum subsp. infantis reveals adaptations for milk utilization within the infant microbiome. Proc Natl Acad Sci U S A 2008;105(48):18964-18969.

16. Milani C, Lugli GA, Duranti S, Turroni F, Mancabelli L et al. Bifidobacteria exhibit social behavior through carbohydrate resource sharing in the gut. Sci Rep 2015;5:15782.

17. Duranti S, Lugli GA, Milani C, James K, Mancabelli L et al. Bifidobacterium bifidum and the infant gut microbiota: an intriguing case of microbe-host co-evolution. Environ Microbiol 2019;21(10):3683-3695.

18. Milani C, Mancabelli L, Lugli GA, Duranti S, Turroni F et al. Exploring Vertical Transmission of Bifidobacteria from Mother to Child. Appl Environ Microbiol 2015;81(20):7078-7087.

19. Martin R, Jimenez E, Heilig H, Fernandez L, Marin ML et al. Isolation of bifidobacteria from breast milk and assessment of the bifidobacterial population by PCR-denaturing gradient gel electrophoresis and quantitative real-time PCR. Appl Environ Microbiol 2009;75(4):965-969.

20. Jost T, Lacroix C, Braegger CP, Rochat F, Chassard C. Vertical mother-neonate transfer of maternal gut bacteria via breastfeeding. Environ Microbiol 2014;16(9):2891-2904.

21. Milani C, Mangifesta M, Mancabelli L, Lugli GA, James K et al. Unveiling bifidobacterial biogeography across the mammalian branch of the tree of life. ISME J 2017;11(12):2834-2847.

22. Mattarelli P, Bonaparte C, Pot B, Biavati B. Proposal to reclassify the three biotypes of Bifidobacterium longum as three subspecies: Bifidobacterium longum subsp. longum subsp. nov., Bifidobacterium longum subsp. infantis comb. nov. and Bifidobacterium longum subsp. suis comb. nov. Int J Syst Evol Microbiol 2008;58(Pt 4):767-772.

23. Matteuzzi D, Crociani F, Zani G, Trovatelli LD. Bifidobacterium suis n. sp.: a new species of the genus Bifidobacterium isolated from pig feces. Z Allg Mikrobiol 1971;11(5):387-395.

24. Odamaki T, Bottacini F, Kato K, Mitsuyama E, Yoshida K et al. Genomic diversity and distribution of Bifidobacterium longum subsp. longum across the human lifespan. Sci Rep 2018;8(1):85.

25. Oki K, Akiyama T, Matsuda K, Gawad A, Makino H et al. Long-term colonization exceeding six years from early infancy of Bifidobacterium longum subsp. longum in human gut. BMC Microbiol 2018;18(1):209.

26. Kitaoka M. Bifidobacterial enzymes involved in the metabolism of human milk oligosaccharides. Adv Nutr 2012;3(3):422S-429S.

27. Martin R, Makino H, Cetinyurek Yavuz A, Ben-Amor K, Roelofs M et al. Early-Life Events, Including Mode of Delivery and Type of Feeding, Siblings and Gender, Shape the Developing Gut Microbiota. PLoS One 2016;11(6):e0158498.

28. Zabel B, Yde CC, Roos P, Marcussen J, Jensen HM et al. Novel Genes and Metabolite Trends in Bifidobacterium longum subsp. infantis Bi-26 Metabolism of Human Milk Oligosaccharide 2'-fucosyllactose. Sci Rep 2019;9(1):7983.

29. Lugli GA, Duranti S, Albert K, Mancabelli L, Napoli S et al. Unveiling Genomic Diversity among Members of the Species Bifidobacterium pseudolongum, a Widely Distributed Gut Commensal of the Animal Kingdom. Appl Environ Microbiol 2019;85(8).

30. Matsuki T, Watanabe K, Tanaka R, Fukuda M, Oyaizu H. Distribution of bifidobacterial species in human intestinal microflora examined with 16S rRNA-gene-targeted species-specific primers. Appl Environ Microbiol 1999;65(10):4506-4512.

31. Ventura M, Meylan V, Zink R. Identification and tracing of Bifidobacterium species by use of enterobacterial repetitive intergenic consensus sequences. Appl Environ Microbiol 2003;69(7):4296-4301.

32. Lugli GA, Milani C, Mancabelli L, van Sinderen D, Ventura M. MEGAnnotator: a user-friendly pipeline for microbial genomes assembly and annotation. FEMS Microbiol Lett 2016;363(7).

33. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19(5):455-477.

34. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 2010;11:119.

35. Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics 2012;28(1):125-126.

36. Zhao Y, Wu J, Yang J, Sun S, Xiao J et al. PGAP: pan-genomes analysis pipeline. Bioinformatics 2012;28(3):416-418.

37. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 2002;30(7):1575-1584.

38. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002;30(14):3059-3066.

39. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007;23(21):2947-2948.

40. Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 2018;9(1):5114.

41. Waack S, Keller O, Asper R, Brodag T, Damm C et al. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC Bioinformatics 2006;7:142.

42. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res 2015;43(Database issue):D298-299.

43. Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 2007;35(Web Server issue):W52-57.

44. Thurston NE, Kerr JC. A nutritional knowledge questionnaire for the elderly. Can J Public Health 1983;74(4):256-260.

45. Carattoli A, Zankari E, Garcia-Fernandez A, Voldby Larsen M, Lund O et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother 2014;58(7):3895-3903.

46. Duranti S, Milani C, Lugli GA, Turroni F, Mancabelli L et al. Insights from genomes of representatives of the human gut commensal Bifidobacterium bifidum. Environ Microbiol 2015;17(7):2515-2531.

47. Lugli GA, Tarracchini C, Alessandri G, Milani C, Mancabelli L et al. Decoding the Genomic Variability among Members of the Bifidobacterium dentium Species. Microorganisms 2020;8(11).

48. Albert K, Rani A, Sela DA. Comparative Pangenomics of the Mammalian Gut Commensal Bifidobacterium longum. Microorganisms 2019;8(1).

49. Freitas AC, Hill JE. Bifidobacteria isolated from vaginal and gut microbiomes are indistinguishable by comparative genomics. PLoS One 2018;13(4):e0196290.

50. Duranti S, Milani C, Lugli GA, Mancabelli L, Turroni F et al. Evaluation of genetic diversity among strains of the human gut commensal Bifidobacterium adolescentis. Sci Rep 2016;6:23971.

51. Lugli GA, Milani C, Turroni F, Duranti S, Mancabelli L et al. Comparative genomic and phylogenomic analyses of the Bifidobacteriaceae family. BMC Genomics 2017;18(1):568.

52. Bottacini F, O'Connell Motherway M, Kuczynski J, O'Connell KJ, Serafini F et al. Comparative genomics of the Bifidobacterium breve taxon. BMC Genomics 2014;15:170.

53. Lugli GA, Milani C, Duranti S, Mancabelli L, Mangifesta M et al. Tracking the Taxonomy of the Genus Bifidobacterium Based on a Phylogenomic Approach. Appl Environ Microbiol 2018;84(4).

54. Tarracchini C, Lugli GA, Mancabelli L, Milani C, Turroni F et al. Assessing the Genomic Variability of Gardnerella vaginalis through Comparative Genomic Analyses: Evolutionary and Ecological Implications. Appl Environ Microbiol 2020;87(1).

55. Fukuda S, Toh H, Hase K, Oshima K, Nakanishi Y et al. Bifidobacteria can protect from enteropathogenic infection through production of acetate. Nature 2011;469(7331):543-547.

56. Prasanna PH, Bell A, Grandison AS, Charalampopoulos D. Emulsifying, rheological and physicochemical properties of exopolysaccharide produced by Bifidobacterium longum subsp.

infantis CCUG 52486 and Bifidobacterium infantis NCIMB 702205. Carbohydr Polym 2012;90(1):533-540.

57. Wei YX, Zhang ZY, Liu C, Zhu YZ, Zhu YQ et al. Complete genome sequence of Bifidobacterium longum JDM301. J Bacteriol 2010;192(15):4076-4077.

58. Koskiniemi S, Sun S, Berg OG, Andersson DI. Selection-driven gene loss in bacteria. Plos Genet 2012;8(6):e1002787.

59. Paul S, Sokurenko EV, Chattopadhyay S. Corrected Genome Annotations Reveal Gene Loss and Antibiotic Resistance as Drivers in the Fitness Evolution of Salmonella enterica Serovar Typhimurium. J Bacteriol 2016;198(23):3152-3161.

60. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev 2005;15(6):589-594.

61. Maglott DR, Katz KS, Sicotte H, Pruitt KD. NCBI's LocusLink and RefSeq. Nucleic Acids Res 2000;28(1):126-128.

62. Jones P, Binns D, Chang HY, Fraser M, Li W et al. InterProScan 5: genome-scale protein function classification. Bioinformatics 2014;30(9):1236-1240.

63. Sarrazin S, Joosten P, Van Gompel L, Luiken REC, Mevius DJ et al. Quantitative and qualitative analysis of antimicrobial usage patterns in 180 selected farrow-to-finish pig farms from nine European countries based on single batch and purchase data. J Antimicrob Chemother 2019;74(3):807-816.

64. Burow E, Rostalski A, Harlizius J, Gangl A, Simoneit C et al. Antibiotic resistance in Escherichia coli from pigs from birth to slaughter and its association with antibiotic treatment. Prev Vet Med 2019;165:52-62.

65. Moller PL, Jorgensen F, Hansen OC, Madsen SM, Stougaard P. Intra- and extracellular beta-galactosidases from Bifidobacterium bifidum and B. infantis: molecular cloning, heterologous expression, and comparative characterization. Appl Environ Microbiol 2001;67(5):2276-2283.

66. Gueimonde M, Salminen S, Isolauri E. Presence of specific antibiotic (tet) resistance genes in infant faecal microbiota. FEMS Immunol Med Microbiol 2006;48(1):21-25.

67. Roberts MC. Update on acquired tetracycline resistance genes. FEMS Microbiol Lett 2005;245(2):195-203.

68. Davani-Davari D, Negahdaripour M, Karimzadeh I, Seifan M, Mohkam M et al. Prebiotics: Definition, Types, Sources, Mechanisms, and Clinical Applications. Foods 2019;8(3).

69. Wicinski M, Sawicka E, Gebalski J, Kubiak K, Malinowski B. Human Milk Oligosaccharides: Health Benefits, Potential Applications in Infant Formulas, and Pharmacology. Nutrients 2020;12(1).

70. Duar RM, Casaburi G, Mitchell RD, Scofield LNC, Ortega Ramirez CA et al. Comparative Genome Analysis of Bifidobacterium longum subsp. infantis Strains Reveals Variation in Human Milk Oligosaccharide Utilization Genes among Commercial Probiotics. Nutrients 2020;12(11).

71. Sela DA, Li Y, Lerno L, Wu S, Marcobal AM et al. An infant-associated bacterial commensal utilizes breast milk sialyloligosaccharides. J Biol Chem 2011;286(14):11909-11918.

72. Intra J, Pavesi G, Horner DS. Phylogenetic analyses suggest multiple changes of substrate specificity within the glycosyl hydrolase 20 family. BMC Evol Biol 2008;8:214.

73. Sakanaka M, Gotoh A, Yoshida K, Odamaki T, Koguchi H et al. Varied Pathways of Infant Gut-Associated Bifidobacterium to Assimilate Human Milk Oligosaccharides: Prevalence of the Gene Set and Its Correlation with Bifidobacteria-Rich Microbiota Formation. Nutrients 2019;12(1).

74. Duranti S, Turroni F, Lugli GA, Milani C, Viappiani A et al. Genomic characterization and transcriptional studies of the starch-utilizing strain Bifidobacterium adolescentis 22L. Appl Environ Microbiol 2014;80(19):6080-6090.

75. Zabel BE, Gerdes S, Evans KC, Nedveck D, Singles SK et al. Strain-specific strategies of 2'-fucosyllactose, 3-fucosyllactose, and difucosyllactose assimilation by Bifidobacterium longum subsp. infantis Bi-26 and ATCC 15697. Sci Rep 2020;10(1):15919.

76. Duranti S, Lugli GA, Mancabelli L, Armanini F, Turroni F et al. Maternal inheritance of bifidobacterial communities and bifidophages in infants through vertical transmission. Microbiome 2017;5(1):66.

77. Garrido D, Ruiz-Moyano S, Lemay DG, Sela DA, German JB et al. Comparative transcriptomics reveals key differences in the response to milk oligosaccharides of infant gut-associated bifidobacteria. Sci Rep 2015;5:13517.

78. Garrido D, Ruiz-Moyano S, Kirmiz N, Davis JC, Totten SM et al. A novel gene cluster allows preferential utilization of fucosylated milk oligosaccharides in Bifidobacterium longum subsp. longum SC596. Sci Rep 2016;6:35045.

79. Kato K, Odamaki T, Mitsuyama E, Sugahara H, Xiao JZ et al. Age-Related Changes in the Composition of Gut Bifidobacterium Species. Curr Microbiol 2017;74(8):987-995.

80. Sela DA, Mills DA. Nursing our microbiota: molecular linkages between bifidobacteria and milk oligosaccharides. Trends Microbiol 2010;18(7):298-307.

81. Kelly SM, Munoz-Munoz J, van Sinderen D. Plant Glycan Metabolism by Bifidobacteria. Front Microbiol 2021;12:609418.

82. Kujawska M, La Rosa SL, Roger LC, Pope PB, Hoyles L et al. Succession of Bifidobacterium longum Strains in Response to a Changing Early Life Nutritional Environment Reveals Dietary Substrate Adaptations. iScience 2020;23(8):101368.

83. Garrido D, Barile D, Mills DA. A molecular basis for bifidobacterial enrichment in the infant gastrointestinal tract. Adv Nutr 2012;3(3):415S-421S.

84. Bolotin E, Hershberg R. Horizontally Acquired Genes Are Often Shared between Closely Related Bacterial Species. Front Microbiol 2017;8:1536.

85. Bonham KS, Wolfe BE, Dutton RJ. Extensive horizontal gene transfer in cheese-associated bacteria. Elife 2017;6.

86. Gyles C, Boerlin P. Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. Vet Pathol 2014;51(2):328-340.

87. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature 2000;405(6784):299-304.

88. Bottacini F, O'Connell Motherway M, Casey E, McDonnell B, Mahony J et al. Discovery of a conjugative megaplasmid in Bifidobacterium breve. Appl Environ Microbiol 2015;81(1):166-176.

89. Ventura M, Turroni F, Foroni E, Duranti S, Giubellini V et al. Analyses of bifidobacterial prophage-like sequences. Antonie Van Leeuwenhoek 2010;98(1):39-50.

90. Milani C, Lugli GA, Duranti S, Turroni F, Bottacini F et al. Genomic encyclopedia of type strains of the genus Bifidobacterium. Appl Environ Microbiol 2014;80(20):6290-6302.

91. Mancabelli L, Milani C, Lugli GA, Turroni F, Cocconi D et al. Identification of universal gut microbial biomarkers of common human intestinal diseases by meta-analysis. FEMS Microbiol Ecol 2017;93(12).

92. Woods LC, Gorrell RJ, Taylor F, Connallon T, Kwok T et al. Horizontal gene transfer potentiates adaptation by reducing selective constraints on the spread of genetic variation. Proc Natl Acad Sci U S A 2020;117(43):26868-26875.

93. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A 2005;102(39):13950-13955.

94. Collins RE, Higgs PG. Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. Mol Biol Evol 2012;29(11):3413-3425.

95. Wackett LP. Lactic acid bacteria: An annotated selection of World Wide Web sites relevant to the topics in microbial biotechnology. Microb Biotechnol 2016;9(4):525-526.

96. McAuliffe O, Ross RP, Hill C. Lantibiotics: structure, biosynthesis and mode of action. FEMS Microbiol Rev 2001;25(3):285-308.

97. Oliveira PH, Touchon M, Rocha EP. Regulation of genetic flux between bacteria by restriction-modification systems. Proc Natl Acad Sci U S A 2016;113(20):5658-5663.

98. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science 2007;315(5819):1709-1712.

99. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E et al. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol 2011;9(6):467-477.

100. Hidalgo-Cantabrana C, Crawley AB, Sanchez B, Barrangou R. Characterization and Exploitation of CRISPR Loci in Bifidobacterium longum. Front Microbiol 2017;8:1851.

Chapter 5

The core genome evolution of *Lactobacillus crispatus* as a driving force for niche competition in the human vaginal tract

Chiara Tarracchini # , Chiara Argentini # , Giulia Alessandri, Gabriele Andrea Lugli, Leonardo Mancabelli, Federico Fontana, Rosaria Anzalone, Alice Viappiani, Francesca Turroni, Marco Ventura, Christian Milani.

The results of this chapter were published in Microbial Biotechnology, 2023 Sep; doi: 10.1111/1751-7915.14305.

#These authors contributed equally.

Abstract

The lower female reproductive tract is notoriously dominated by *Lactobacillus* species, among which *Lactobacillus crispatus* emerges for its protective and health-promoting activities.

Although previous comparative genome analyses highlighted genetic and phenotypic diversity within the *L. crispatus* species, most studies have focused on the presence/absence of accessory genes. Here, we investigated the variation at the single nucleotide level within protein-encoding genes shared across a human-derived *L. crispatus* strain selection, which includes 200 currently available human-derived *L. crispatus* genomes as well as 41 chromosome sequences of such taxon that have been decoded in the framework of this study. Such data clearly pointed out the presence of intra-species micro-diversities that could have evolutionary significance contributing to phenotypical diversification by affecting protein domains. Specifically, two single nucleotide variations in the type II pullulanase gene sequence led to specific amino acid substitutions, possibly explaining the substantial differences in the growth performances and competition abilities observed in a multi-strain bioreactor culture simulating the vaginal environment.

Accordingly, *L. crispatus* strains display different growth performances suggesting that the colonization and stable persistence in the female reproductive tract between the members of this taxon is highly variable.

For Supplementary Materials see the article published in Microbial Biotechnology

INTRODUCTION

Bacteria evolved over millions of years to colonize different districts of the human body, e.g., skin, pulmonary, gastrointestinal, and vaginal tracts, giving rise to complex and dynamic populations of microorganisms engaged in close relationships with the human host, referred to as the microbiota (1). In particular, the gut microbiota, with its vastity of microbial genera and species, has attracted increasing interest in the last decades for its ability to impact several aspects of human health, development, and systemic physiology from infancy to adulthood (2–7). In contrast, the vaginal microbiome is typically manifested by a low degree of (bio)diversity and is commonly dominated by members of the *Lactobacillus* genus, such as *Lactobacillus iners*, *Lactobacillus gasseri*, *Lactobacillus jensenii*, and *Lactobacillus crispatus*. This latter is regarded as the primary determinant of vaginal health (8,9). Indeed, in healthy cervicovaginal microbiota, *L. crispatus* species prevails, producing D- and L-lactic acid, hydrogen peroxide, and bacteriocins, which prevent the overgrowth of possible pathogens, hence preventing upper genital tract infections in the host (10,11). For such reason, probiotic supplements based on *L. crispatus* are widely used as vehicles of health-promoting strains in the vaginal environment (12–14).

Recently, the evolution of the genome sequences of *L. crispatus* species has been studied in relation to its adaptation to the human vaginal niche, underlining strain-dependent efficiency to grow on glycogen as well as to inhibit pathogens (15–20). Moreover, besides the human vaginal tract, *L. crispatus* have also been identified and isolated from various (sub)niches, ranging from healthy poultry gut to various districts of the human body, including the oral cavity, rectum, and urinary tract, highlighting within-species genetic diversity, and variegated metabolic capabilities (21–24). Taken together, this evidence suggests the presence of distinct evolutive trajectories underlying the observed phenotypic diversification within this species.

However, comparative genomic analyses involving chromosome sequences of *L. crispatus* species have often focused on the relationship between the presence/absence of accessory genes and a particular phenotype (21,24), overshadowing the importance of mutations to within-species evolution (25–27).

In this framework, the aim of this study is to evaluate genome sequence variations at the single nucleotide level within protein-encoding genes shared across non-identical *L. crispatus* chromosomes, providing a close-up view of genetic (micro)diversity, which can contribute significantly to strain diversification within this species. In addition, to investigate the possible implications of the identified genetic differences in the *L. crispatus* intraspecies competition within the vaginal microbiota, we performed *in vitro* experiments consisting of carbohydrate-growth assays involving *L. crispatus* multi-strain co-cultivation in a bioreactor simulating the vaginal tract.

Our findings revealed inter-strain genotypic variation and phenotypic differences between *L. crispatus* strains, highlighting distinct evolutionary developments that may provide this species with differential abilities to long persist and predominate in the human vaginal tract.

RESULTS AND DISCUSSION

Identification of representative *Lactobacillus crispatus* genomes

To investigate the genomic differences between human *L. crispatus* strains, an extensive comparative genome analysis was performed on *L. crispatus* genomes recovered from human specimens of healthy donors, including fecal, vaginal, saliva, and urine samples (Supplementary Table S1). Specifically, seven *L. crispatus* strains were obtained from international bacteria culture collection (Table 1), and their genomes were sequenced along with those of 34 strains isolated from the human vaginal tract in the context of a previous study (Table 1). Additionally, with the aim of expanding the overview of the genetic variability of this taxon, 200 genome sequences (complete and draft) of *L. crispatus* strains isolated from human biological samples were selected from public repositories (Table S1). Following dereplication aimed at removing the genomic redundancy by grouping essentially identical genomes (using dRep tool, version 2.2.0, with average nucleotide identity > 99 %, [https://drep.readthedocs.io/en/latest/choosing_parameters.html]), 22 *L. crispatus* chromosomes with average completeness of 98.97 % ± 0.14 % were retained as representatives of the sequence variation observed in our genome repertoire and therefore used for comparative analysis (Supplementary Table S2).

The general features of the 22 representative *L. crispatus* genomes are reported in Table 1 and include an average of 2,105 ± 179 predicted Coding Sequences (CDSs) per chromosome (ranging from 2,632 to 1,865), with an average genome length of 2.20 ± 0.18 Mbp.

Table 1. Genome features of the 22 representative *L. crispatus* genomes.

Assembly ID	Strain name	Genome size Mbp	CDS number	Genome Completeness (%)	Isolation source
GCF_000162255.1	125-2-CHN	2.30525	2,032	99.03	human vagina
GCF_000162315.1	MV-3A-US	2.43708	2,252	98.38	human vagina
GCF_002861805.1	UMB0824	2.17405	2,061	99.03	human urine
GCF_002861815.1	UMB0085	2.17506	2,081	99.03	human urine
GCF_009857395.1	Indica2	2.20949	2,028	99.03	human vagina
GCF_007713895.1	NCK1350	2.04734	1,932	99.03	human stool
GCF_013456995.1	B4	2.03959	1,902	99.03	human stool
GCF_000160515.1	JV-V01	2.2172	1,992	98.03	human vagina
GCF_014654865.1	BC5	2.06419	1,901	99.03	human vagina
GCF_015669875.1	D31t1	2.2782	2,120	99.03	human stool
GCF_018987235.1	ATCC 33820	2.23909	2,020	99.03	human saliva
GCF_020042005.1	Lc1700	2.81896	2,632	98.86	human vagina
GCF_021278925.1	lc83	2.30843	2,112	98.9	human vagina
GCF_025194085.1	CIRM-BIA 2111	2.00737	1,865	99.03	human stool
GCF_025194045.1	CIRM-BIA 2233	2.24513	2,127	99.03	human vagina
This study	LMG11440	2.032412	2,087	98.86	human vagina
This study	LMG12005	2.019682	2,014	98.94	human vagina
This study	LMG18189	2.094399	2,079	99.03	human saliva
This study	LMG11415	2.030901	2,004	99.03	human saliva
This study	LMG18200	2.208098	2,182	99.03	human stool
This study	LB93	2.202822	2,340	99.03	human vagina
This study	LB97	2.263389	2,422	99.03	human vagina

Intra-species genetic variability within the *Lactobacillus* genus

To investigate the level of genomic diversity among *L. crispatus* strains compared with other species of the *Lactobacillus* genus, we selected publicly accessible chromosomes belonging to seven different *Lactobacillus* species known to inhabit various human body sites. Notably, for a robust comparison with the dereplicated 22 representative *L. crispatus* genomes, we focused on *Lactobacillus* species for which at least 20 independent conspecific genomes with ANI values between 95% and 98% were retained after accounting for genome completeness > 95 % (Supplementary Table S3). Accordingly, 499 *Lactobacillus* chromosomes were collected and

combined with the 22 representative genomes of *L. crispatus* for pangenome analysis, which led to the identification of 159 core genes, defined as the set of gene families (clusters of orthologous groups [COGs]) shared by each *Lactobacillus* chromosome tested (Figure 1a).

Exploiting this set of 159 core genes, the level of variability at the single nucleotide level was evaluated individually within each *Lactobacillus* species. In detail, sequences homologous to the 159 core genes (corresponding to an average of 46,760 ± 987 nucleotides) were recovered from each collected genome and compared between strains belonging to the same *Lactobacillus* species, eventually recording Single Nucleotide Polymorphisms (SNPs) at each nucleotide position.

Considering the nucleotide variations with an occurrence rate above 20% to exclude sequencing errors, the eight inspected *Lactobacillus* lineages exhibited a total number of intra-species SNPs per Mbp ranging from 20.6 to 442.3, with *L. johnsonii* and *L. acidophilus* showing the lowest and the highest number of SNPs, respectively (Figure 1a). Thus, these data highlighted how, on average, the various species of the *Lactobacillus* genus display different levels of intra-species genetic diversity, which has the potential to translate into intra-species phenotypic variability.

Remarkably, the *L. crispatus* core gene set returned an average of 364.7 SNPs per Mbp, emerging among the species with the higher genetic variation, even compared with other notorious *Lactobacillus* species inhabiting the human vaginal tract, such as *L. gasseri*, *L. iners*, and *L. jensenii* (Figure 1a).

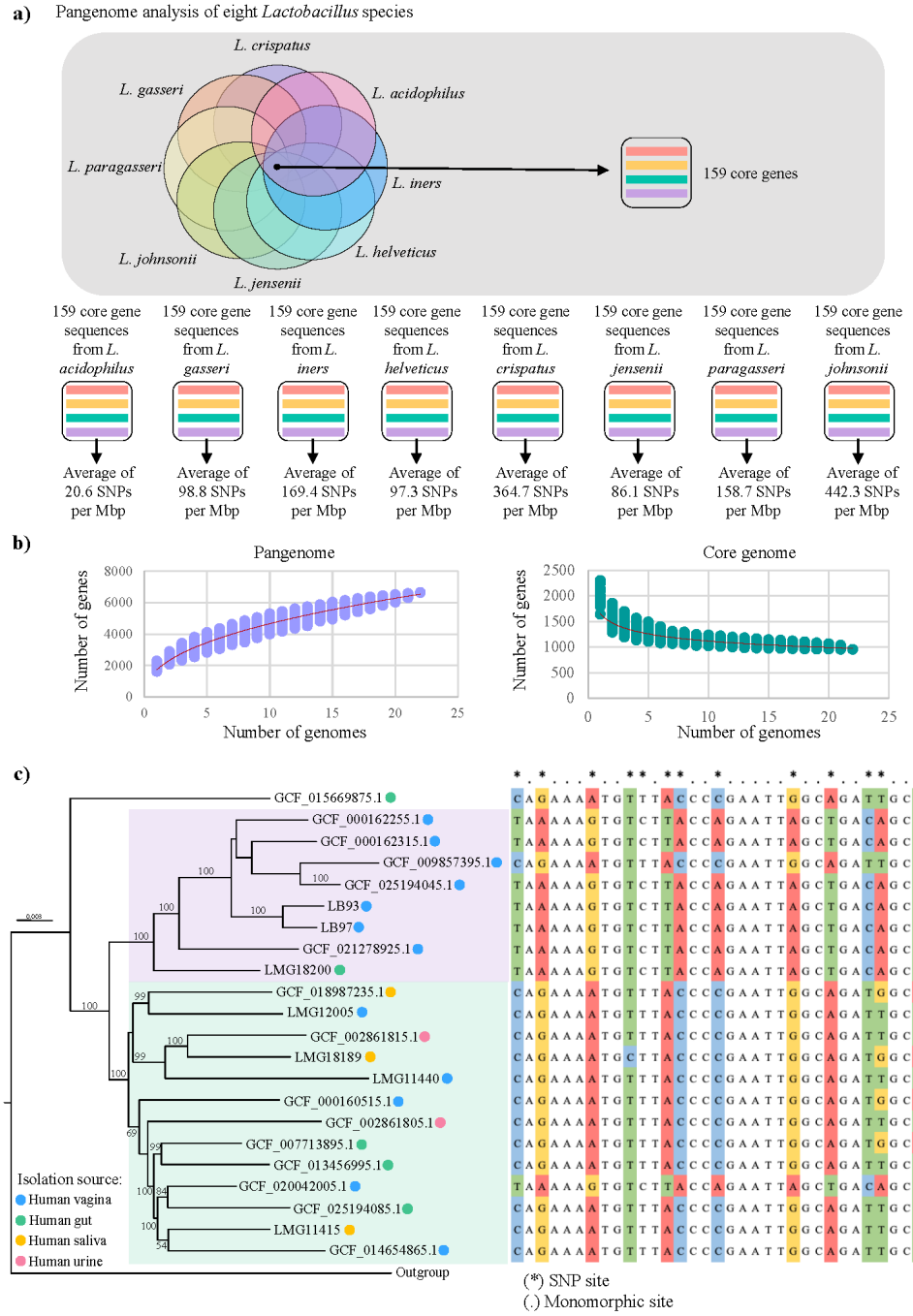


Figure 1. Comparative analysis between different *Lactobacillus* species and between 22 non-identical *L. crispatus* strains. In panel **(a)**, Venn diagram shows the eight *Lactobacillus* species sharing the 159 genes used to measure the magnitude of intra-species genetic diversity. Below, the species-specific number of SNPs identified within the common protein-encoding genes is reported for each of the considered *Lactobacillus* species. Panel **(b)** depicts the *L. crispatus* pan- and core-genome size. The number of discovered genes (vertical axe)

is plotted as a function of the number of sequentially added genomes (horizontal axe). Panel **(c)** shows the phylogenomic tree based on the concatenated 959 core genes shared among the 22 non-identical *L. crispatus* genomes. The tree was constructed by the neighbor-joining method. Bootstrap percentages based on 1,000 replicates above 50 are shown at node points. For each strain, the isolation source is highlighted with a colored circle. On the right, an aligned portion of the *L. crispatus* core genome exemplifies the relationships between phylogenomic clusters and SNP patterns. In the top row, nucleotide positions showing variants are highlighted with an asterisk, while a dot highlights non-variant sites.

In particular, among *L. crispatus* members, most SNPs were found within gene sequences coding for transmembrane transport mechanisms (26 %), followed by genes involved in the biosynthesis of extracellular protein components (14.5 %) and carbohydrate metabolism (10.4 %). In contrast, within more niche-specialized *Lactobacillus* species, e.g., *L. gasseri*, the genes with higher SNPs were involved DNA-related processes (21.4 %) and protein-protein interaction (12 %).

Pan- and Core-genome analysis of the *L. crispatus* species

Chromosome sequences of the 22 non-identical *L. crispatus* strains were submitted to gene re-annotation and subsequently analyzed from a pangenome perspective, providing information on the ubiquitous genetic backbone of conspecific chromosomes and the intra-species genetic diversity (28). In total, the pangenome of *L. crispatus* includes 6,512 COGs, whose accumulation curve, depicting the expansion of the pangenome as a function of the number of genomes included, is still far from being saturated. Thus, indicating that *L. crispatus* species is characterized by an open pangenome where the total gene pool obtainable for this species has not yet been fully disclosed (Figure 1b). Moreover, we determined the current *L. crispatus* core genome to be comprised of 959 COGs that were conserved across all the 22 analyzed strains (15 % of the pangenome), while an average of 157.3 ± 54.3 genes per genome were associated with only one strain.

Based on the core gene sequences obtained from the 22 non-redundant *L. crispatus* genomes, a phylogenetic tree was constructed to evaluate the evolution of the species (Figure 1c). According to the clustering relationship, the 22 strains assessed were divided into two main clusters, one of which was intriguingly composed only of *L. crispatus* strains isolated from the female reproductive tract (Figure 1c, violet shadows). Moreover, this phylogenetic tree also displayed a second phylogenetic cluster of strains isolated from different human body districts, encompassing vagina, gut, saliva, and urine (Figure 1c, green shadow). Notably, this mixed group may include a few strains that can survive/colonize in closely related niches, like different human body districts.

To investigate the intra-species genomic variability of *L. crispatus* taxon, we measured the genetic diversity at the single-nucleotide level by comparing the whole genome sequences (wgSNPs) and the corresponding core genome (cgSNPs). Specifically, from the core genome-based phylogenomic tree (Figure 1), we selected the *L. crispatus* strain placed at the deepest split, i.e., the most divergent chromosome (RefSeq genome assembly GCF_015669875), which was used as a reference sequence to compute pairwise alignments and SNPs extraction. Overall, the 22 *L. crispatus* strains showed an average of 28,811 wgSNPs (representing about 2 % of the genome sequence), most of which (about 90 %) resided within coding sequences. For cgSNPs evaluation, each of the 22 homologous nucleotide sequences obtained by concatenating the 959 COGs shared among all the non-identical *L. crispatus* strains (corresponding to an average of 904,903 bases) were examined for sequence variations against the concatenated (core) gene set of the reference *L. crispatus* assembly. This procedure resulted in the identification of an average of 15,007 cgSNPs (representing a variation rate of 1 for every 56 nucleotides), that were lowered when compared with previous analyses of polymorphic sites within

clinically relevant organisms such as *Pseudomonas aeruginosa* and *Escherichia coli* (showing 159,609 SNPs within the concatenated core genes of 3,629,979 bp and 266,969 SNPs within a core genome of 2,159,296 bp, respectively) (29,30). Albeit the core genome resulted rather conserved within these 22 *L. crispatus* strains, it might be worth mentioning that the observed micro-diversity lies within DNA sequences that code for proteins. Therefore, it could have evolutionary importance contributing to intra-species diversity by affecting protein domains. Indeed, the nucleotide sequence variation among the *L. crispatus* core genome was not randomly distributed, but phylogenetically co-clustered strains showed common patterns of SNP profiles (Figure 1c), thus indicating that the observed SNPs are representing evolutionary trajectories and not mere random mutations or sequencing errors.

Exploration of the micro-diversity in the ubiquitous features of *L. crispatus* genome and identification of fast evolving genes

With the aim of defining whether and which categories of genes are more concerned by a rapid sequence evolution, we calculated the level of polymorphisms for each protein-coding gene constituting the *L. crispatus* core genome. Like what has been performed above, the homologous gene sequences from the *L. crispatus* GCF_015669875 were used as reference in pairwise comparisons of each individual core gene. Accordingly, the number of SNPs resulting from the average of all alignment pairs ranged from zero to 347.45 ± 150.30 per gene (Supplementary Table S4).

Among the genes with the lower average number of SNP sites (lower than 5.8, corresponding to the data below the 25th percentile, Figure 2a), we identified protein-coding sequences involved in putative housekeeping functions, including ribosome assembly and function, central glycolysis regulation, as well as DNA replication and cell division (Supplementary Table S4).

In contrast, by considering the data above the third quartile (Q3+1.5·Inter-Quartile Range, Figure 2a), we identified 52 genes with the highest average number of SNP sites (ranging from 347.45 ± 150.30 to 90.25 ± 28.76), i.e., the most highly variable genes (HVGs), which therefore represented the set of genes that have been likely under the strongest selection pressure (Figure 2b, Supplementary Table S4). Interestingly, among the HVGs, we identified mainly genes involved in the biosynthesis and rearrangement of cell wall components, such as lipoteichoic acids and peptidoglycan, as well as transmembrane transport of a variety of substrates, including carbohydrates and micronutrients (Figure 2b, Supplementary Table S4).

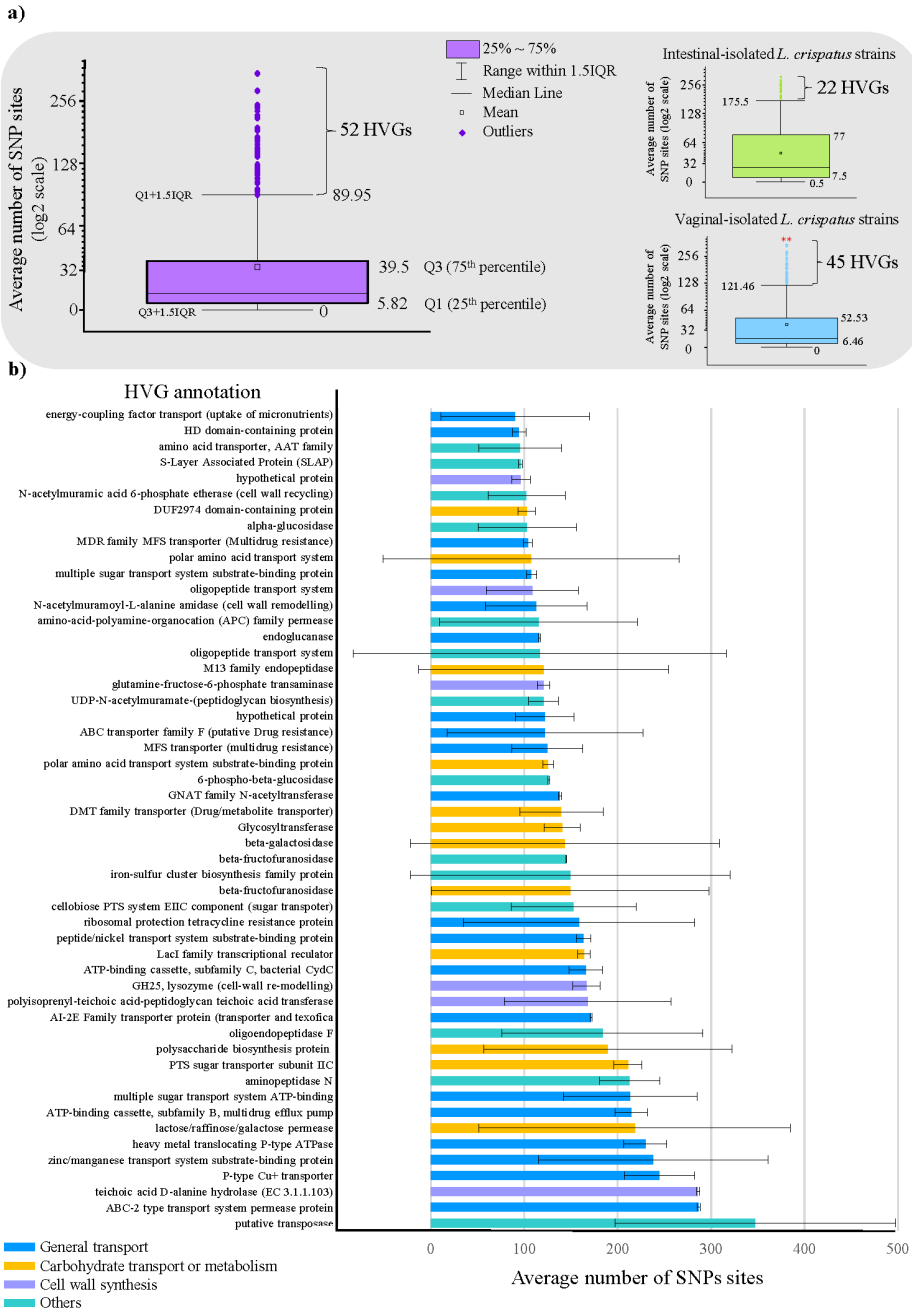


Figure 2. Identification of the 52 HVGs. In panel **(a)**, Box-Whisker plot was used to represent the gene distribution based on the number of SNP sites obtained by comparing the nucleotide sequence of every 959 protein-coding genes shared among all the 22 *L. crispatus* chromosomes. For each gene, the number of SNP sites was expressed as the average of all the pairwise comparisons against the reference sequence (homologous gene sequence of *L. crispatus* GCF_015669875). The Q3+1.5IQR was used as a cut-off to select the 52 HVGs. Panel **(b)** reports the functional annotation and the number of SNP sites of each HVG.

The presence of such micro-diversity in proteins directly mediating interactions with the environment likely reflects adaptive mechanisms to the changing biotic and abiotic components, thereby leading to possible different competitive abilities and (sub-)niche specialization. Indeed, the intra-species heterogeneity observed in the *L. crispatus* core genes emerged less marked when the 22 *L. crispatus* strains were compared based on their ecological niche, thus showing greater gene sequence homogeneity among genomes that share the same environment (Figure 2a). However, vaginal-derived *L. crispatus* strains showed a significantly higher number of HVGs than those isolated from the human gut (Mann-Whitney test, *p*-value = 0.007), indicating that the vaginal environment exerts crucial ecological forces driving the *L. crispatus* genome evolution.

A new phylogenomic tree, representing the evolutionary outcomes determined by mutational hotspots within the *L. crispatus* species, was generated employing the nucleotide sequences of the 52 HVGs (Figure 3). Specifically, this tree was composed of three main clusters, where not all the strains maintained the same distribution compared to the original phylogenomic tree based on the whole core genome (Figure 3). Indeed, the HVG-based tree better distinguished among strains from closely related niches, highlighting that the evolution of the HVGs does not strictly follow the overall strain speciation, probably reflecting a relatively recent adaptation to specific environmental stimuli, such as multiple human body site colonization or inter-strain niche competition. Accordingly, based on the picture emerging from the HVG-derived phylogenetic distribution, we selected four highly divergent *L. crispatus* strains, i.e., LB97, LMG11440, LMG18200, and LMG11415, that were used for *in vitro* phenotypical assays aimed at investigating the link between evolutionary trajectories, grow performances, and competitive abilities.

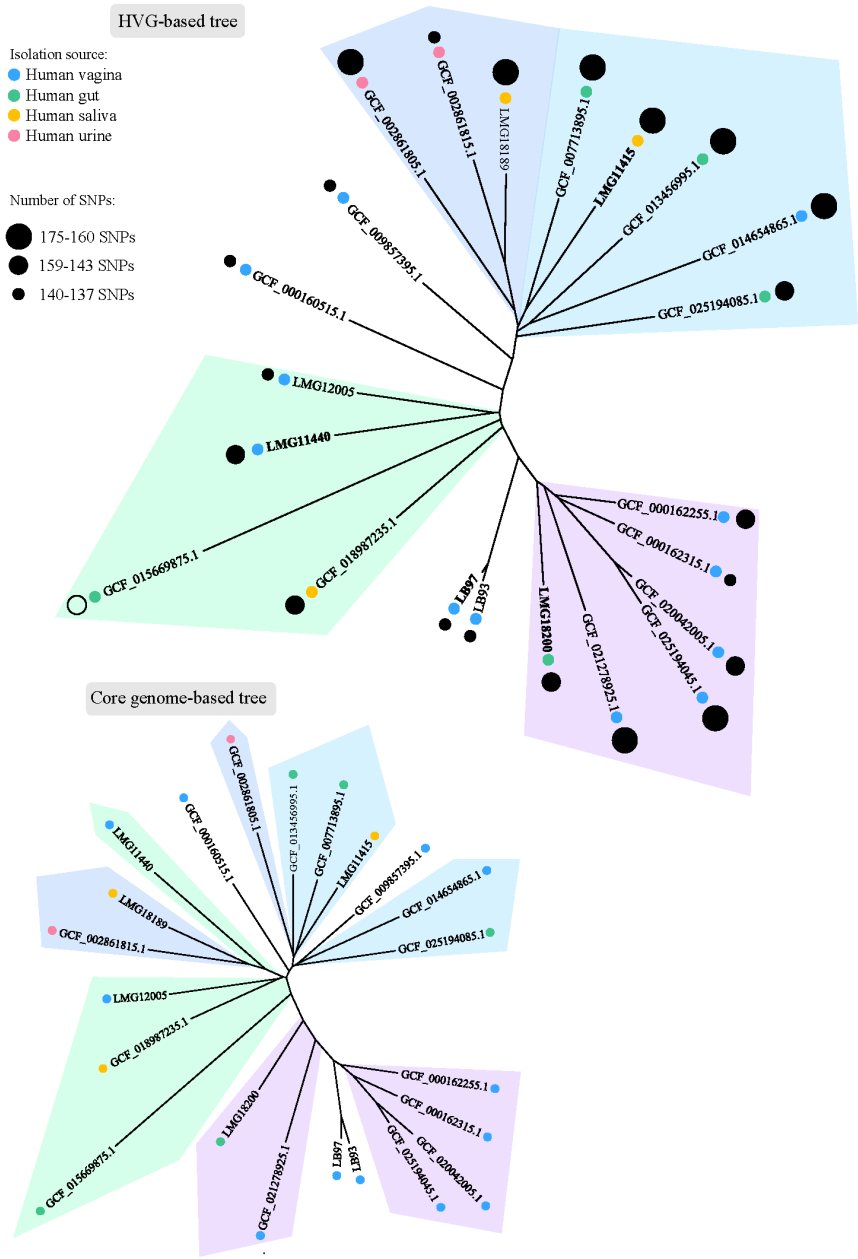


Figure 3. Phylogenetic analysis based on the 52 HVG sequences. Proteomic tree based on concatenating the 52 protein-encoding core genes identified as highly variable across the 22 non-identical *L. crispatus* genomes. Phylogenetic groups are highlighted in different colors. For comparison, the phylogenomic tree resulting from the whole core genome (presented in Figure 1) is visualized using the redial layout. For each strain, the colored circle represents the isolation source, while the diameter of the black circle is proportional to the number of SNPs identified within the core genome using the GCF_015669875.1 genome sequence as reference.

In vitro evaluation of *L. crispatus* growth performances on selected carbohydrate sources in an *in vitro* human vaginal model

To investigate how the high level of genetic heterogeneity identified within *L. crispatus* could influence the respective growth abilities, the four selected *L. crispatus* strains (LB97, LMG11440, LMG18200, and LMG11415, whose genome diversity corresponded to ANI pairwise values < 98 %, Supplementary Table S2) were cultivated on six different carbohydrate sources including glycogen, which is the primary bacterial nutritional source in the vaginal lumen (31,32), along with other glycogen-like α-glucans which may also represent a substrate for the bacterial enzymatic arsenal involved in carbohydrates breakdown of the vaginal environment (Figure 4a, Supplementary Table S5). The optical density (OD) was registered after 48h of anaerobic growth, and the growth on MRS was used as control condition (Supplementary Table S5). Upon one-way ANOVA test with Bonferroni correction (cut-off p-value < 0.05), the comparative growth assay showed widespread statistically significant differences across the four *L. crispatus* strains. In particular, *L. crispatus* LB97, isolated from the human vagina, showed greater growth performances on most of the carbohydrates tested, including glycogen (final OD > 1.2; all Bonferroni-corrected p-values < 0.05), thus demonstrating a metabolic specialization consistent with its isolation niche.

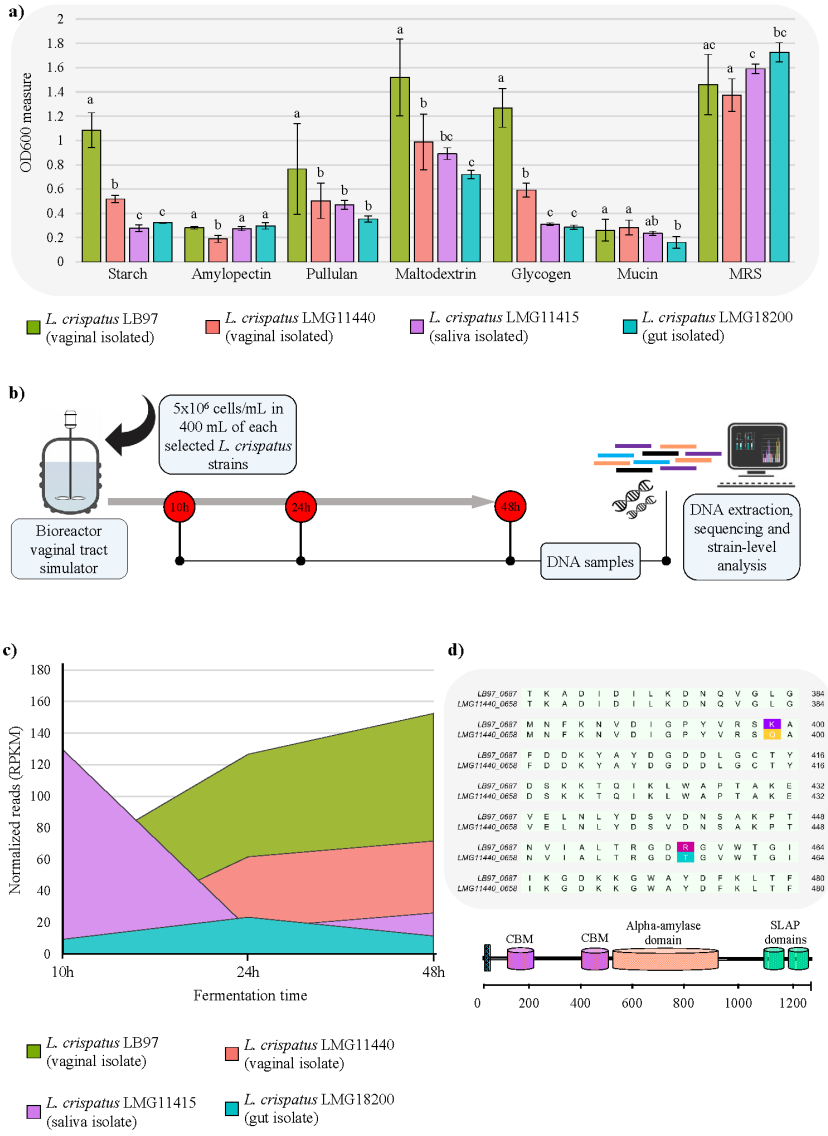


Figure 4. Differential growth and competitive abilities between *L. crispatus* strains. Panel **(a)** shows the optical density (OD) registered after 48h of anaerobic growth in different nutritive substrates. Panel **(b)** illustrates the design of the bioreactor-based experiment simulating the vaginal environment. In panel **(c)**, the bar chart reports the quantification of the metagenomic reads (using average RPKM measures) mapping marker genes unique to each *L. crispatus* strain throughout the 48 hours of growth in the bioreactor.
The standard deviations are plotted as error bars. Different lowercase letters indicate significant differences at p-value < 0.05 according to the Bonferroni test.
In panel **(d)**, alignment of partial amino acid sequences corresponding to the type II pullulanase genes of *L. crispatus* LB97 and LMG11440 strains genes highlights two amino acid substitutions Gln (Q) to Lys (K) and Arg (R) to Thr (T).

Conversely, the gut-derived *L. crispatus* LMG18200 exhibited the lowest growth when glycogen, starch, and pullulan were used as the unique carbon source (all final OD measures ~ 0.3; all Bonferroni-corrected *p*-values < 0.05), suggesting the incapability of this strain to metabolize long-chain α-glucans (Figure 4a, Supplementary Tables S5, S6). Moreover, all strains appeared nearly equally limited in mucin utilization (all final OD < 0.3; Bonferroni-corrected *p*-values > 0.05) (Figure 4a, Supplementary Tables S5, S6).

Furthermore, to determine the reciprocal competitive ability of the different *L. crispatus* strains, the growth performances of the four selected strains were evaluated through a co-cultivation experiment involving a bioreactor model simulating the nutritional and chemical-physical conditions of the vaginal environment (24) (Figure 4b). The proliferation trend of each strain was followed for 48h by mapping the sequenced metagenomic reads at multiple time points against a set of strain-specific marker genes (Supplementary Table S7, Supplementary Table S8). In accordance with what was observed in the carbohydrate grow assay, the vaginal isolate LB97 showed a notable proliferation ability, over dominating the four-strain *L. crispatus* community at every co-cultivation time-point (Supplementary Table S8, all Bonferroni-corrected *p*-values < 0.05) (Figure 4c).

In contrast, the strains LMG11415 (isolated from the human saliva) and LMG18200 (isolated from the human intestine) were clearly overwhelmed (Figure 4c). Consistently, close examination of the genomes of these four *L. crispatus* strains revealed that these latter were lacking in any glycogen-degrading encoding gene, while the proliferating vaginal-derived LB97 and LMG11440 strains carried a type II pullulanase acting on both α-1,6- and α-1,4- glycosidic bonds, which therefore achieves complete glycogen degradation, as reported in recent studies (17,32). Nevertheless, LB97 and LMG11440 substantially differ in their proliferation

capabilities (Supplementary Table S8, Bonferroni-corrected p-values < 0.05) (Figure 4c).

Accordingly, the metabolic potential of the gene sequences identified in the comparative genome analysis as associated uniquely with the *L. crispatus* LB97 was investigated using the MetaCyc database (33). The results revealed that the predicted proteome of this strain is characterized by the presence of protein-encoding genes involved in the uptake and metabolism of galactitol and polyamines, as well as a locus encoding proteins dedicated to the ascorbate degradation, which can contribute to the maintenance of the host's vaginal health (34) (Supplementary Table S9). Moreover, the unique gene repertoire of the LB97 strain also contained a gene encoding for a mucin-binding protein (MucBP domain), thus corroborating for this strain the hypothesis of a host mucin role as adhesion site rather than a carbon source, as previously reported in the literature (29–31) and confirmed by the growth assay described above (Figure 4a). However, our functional investigation did not detect genes possibly involved in the metabolism of the nutritional sources constituting the used glycogen-based culture medium (Supplementary Table S9).

Interestingly, when SNPs were calculated between these latter two *L. crispatus* genome sequences, a total of 27,906 nucleotide positions showed differences, of which 8,238 (29 % of the total SNPs) corresponded to amino acid replacements, thus resulting in strain-specific intragenic variants, which can contribute to generating phenotypic differences (Supplementary Table S10). Moreover, alignment of their type II pullulanase gene revealed four variations at single nucleotide level (Supplementary Table S11, Supplementary Figure S1), two of which resulted in amino acid substitutions (Figure 4d, Figure S1). In more detail, these non-synonymous SNPs lie within the protein Carbohydrate Binding Module (CBM) (Figure 4d), with possible repercussions on the efficiency of the protein binding to

its substrate, as also evidenced by the 3D protein structure prediction (Supplementary Figure S2).

Overall, the finding of isolate-specific intragenic SNPs, and particularly those within the pullulanase-encoding gene, possibly explain the growth and competitiveness differences observed between the vaginal-isolated *L. crispatus* LB97 and LMG11440 strains cultivated on the simulated vaginal medium.

CONCLUSIONS

In this study, a comparative genome analysis involving 41 newly decoded human *L. crispatus* genomes coupled with 200 publicly available genome sequences from this species allowed us to deeply investigate the *L. crispatus* core gene evolution by connecting data from single nucleotide variations, phylogenomic reconstructions, and *in vitro* experiments.

Compared with other *Lactobacillus* species, including those inhabiting the human vaginal tract, i.e., *L. iners, L. gasseri*, and *L. jensenii*, a higher level of sequence variation at the single nucleotide level was observed within the gene pool shared among the inspected *L. crispatus* strains, thus highlighting a within-species diversity driven by conserved genes evolution.

Interestingly, the genetic heterogeneity observed within the *L. crispatus* species appears to be reflected at the phenotypic level. In fact, when different *L. crispatus* strains were co-cultivated in a bioreactor-based model simulating the vaginal environment, substantial differences were noted in the colonization and competition efficiency.

Although members of the *L. crispatus* species was previously thought to utilize the glycogen hydrolysis products generated in the vaginal environment by the human α-amylase (31,35), recent evidence showed that members of this taxon produce the enzyme to independently degrade glycogen, annotated as a type II pullulanase

(17,32). In this context, while the absence of this gene was noted for those *L. crispatus* strains unable to stably proliferate on glycogen under *in vitro* conditions, we identified two amino acid substitutions within the type II pullulanase carbohydrate-binding module arising from non-synonymous single nucleotide polymorphisms (SNPs) which could explain the different proliferation and dominance abilities observed *in vitro* for the *L. crispatus* strains investigated in this study.

Remarkably, while the strain-specific accessory genetic content has been historically pointed out as one of the main sources of variability resulting from intra-species evolution, data collected in the framework of this study revealed that the evolution of the core genome could contribute to generate marked strain-specific phenotypic traits. Thus, understanding this evolutionary driving force could be relevant for unraveling strain-specific capabilities to successfully dominate the female reproductive tract and, ultimately, selecting suitable *L. crispatus* strains that could be applied for novel bacterial therapy strategies.

MATERIAL AND METHODS

Isolation of *L. crispatus* strains and retrieval of publicly available genome sequences

Candidate *Lactobacillus* strains were obtained from an isolation effort performed in a framework a previous study.

Identification of newly isolated *L. crispatus* strains was achieved through the Matrix Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS) Biotyper Sirius (Bruker, Germany) using the manufacturer's software FlexControl and the MALDI-Biotyper software (MBT). In detail, a single bacterial colony grown on MRS agar was transferred onto a spot of the MSP 96 target polished

steel BC MALDI target plate (Bruker, Germany). Subsequently, the bacterial sample was overlaid with 1 μL of matrix solution containing 10 mg/mL HCCA (a-cyano-4-hydroxycinnamic acid, Sigma-Aldrich, Poland) resolved in 50% acetonitrile (Carlo Erba, Italy) and 2.5% TFA (trifluoro-acetic acid, Carlo Erba, Italy) and air-dried (36,37). The MALDI target plate was then introduced into the spectrometer for automated measurement and data interpretation. The mass spectra were processed with the MALDI Biotyper 3.0 software package (Bruker, Germany) containing reference spectra, including different lactobacilli species. According to the criteria recommended by the manufacturer, a score of $\geq$2.000 indicates a significant similarity between the obtained spectrum and the database entry. Each sample was analyzed in duplicate (2 spots for each sample).

The default parameter settings are as follows: positive linear mode, laser frequency 200Hz, ion source 1 = 19.84 kV, ion source 2 = 18.07 kV, Bruker's MBT_FC and MBT_AutoX methods, mass range: 2000–20000Da. Moreover, before analysis, calibration was performed with a bacterial test standard (Bruker, Germany) containing an extract of *Escherichia coli* DH5 alpha.

A total of 34 *L. crispatus* strains were identified and taken forward for whole genome sequencing. In addition, seven *L. crispatus* strains isolated from human biological samples were purchased from international bacterial collections. To perform chromosomal DNA extraction, the 41 *L. crispatus* strains were cultivated in MRS broth supplemented with 0.05% (wt/vol) L-cysteine hydrochloride in an anaerobic atmosphere at 37°C (2.99% [vol/vol] H_2, 17.01% [vol/vol] CO_2, and 80% [vol/vol] N_2) for 12 h. Subsequently, cells from 10 ml of the culture were harvested by centrifugation at 6,000 rpm for 8 min, and the obtained cell pellet was used for DNA extraction using the GenElute bacterial genomic DNA kit (Sigma-Aldrich) following the manufacturer's guide.

Furthermore, 200 *L. crispatus* genome assemblies with completeness > 95% derived from strains isolated from biological material of human subjects were retrieved from the NCBI database.

Genome Sequencing, Assembly, and Annotation

The DNA extracted from the 41 *L. crispatus* strains was subjected to whole-genome sequencing using MiSeq (Illumina, UK) at GenProbio srl, Parma, Italy (www.genprobio.com) according to the supplier's protocol (Illumina, UK). Individual genome libraries were generated using the Nextera XT preparation kit and loaded into a 600-cycle (250-bp paired-ends) flow cell version 3 (Illumina). Raw DNA sequence reads (fastq files) obtained from genome sequencing were assembled using the MEGAnnotator pipeline (38). Briefly, SPAdes software was used for *de novo* assembly of the genome sequences with the pipeline option "--carefull" and a list of k-mer sizes of 21; 33; 55; 77; 99; 127 (39) and protein-encoding genes were predicted for contig greater than 1,000 bp using Prodigal (40). Functionally annotation of the predicted genes was achieved through RAPSearch2 (cut-off E value, 1×10^{-5}; minimum alignment length, 20 amino acids) (41) and hidden Markov model profile (HMM) searches (cut-off E value, 1 × 10−10) (http://hmmer.org/) performed against the NCBI nr database and the manually curated Pfam-A database, respectively. Moreover, tRNA genes were determined using tRNAscan-SE version 1.4 (42), and rRNA loci were identified with RNAmmer version 1.2 (43).

Furthermore, to obtain comparable quality standards for the analyzed genomes, all the 200 *L. crispatus* genomes retrieved from the NCBI database were re-annotated employing the same approach based on MEGAnnotator pipeline used for the 41 *L. crispatus* genomes decoded in the current study.

Pangenome analyses and phylogenomic tree reconstruction

All pangenome calculations were performed using PGAP [PanGenomes Analysis Pipeline, (44)] as described previously (45,46). In detail, orthologous protein sequences were identified in genome sequences using BLAST analysis (cut-off E-value = 1×10^{-5}; 50% identity over at least 80% of sequence coverage) and then organized into functional Clusters of Orthologous Groups (COGs) through the MCL algorithm (graph-based Markov clustering algorithm) using the gene family (GF) method. Pangenome profiles were produced through an optimized procedure integrated into the PGAP software, based on a presence/absence matrix including all COGs identified in the given genomes. The concatenated protein sequences of core genes were aligned using Mafft v7.453 (47) and then employed to build correspondent phylogenomic trees through the neighbor-joining method in ClustalW version 2.1. Visual core genome-based phylogenomic trees were developed using FigTree software (http://tree.bio.ed.ac.uk/software/figtree/).

Single-Nucleotide Polymorphism identification

The species-specific level of polymorphisms within the *Lactobacillus* genus was assessed exploiting the identified core gene set shared between *L. crispatus,* and seven different *Lactobacillus* species. In detail, 159 core-shaping genes were concatenated and aligned using the multiple genome aligner Mafft v7.453 (47). Nucleotide variants at each sequence position were then extracted through the SNP-sites program (version 2.5.1) (48). Assuming that, unlike sequencing errors, real genetic variants should be observed in a quite number of independent genomes assembly, we considered only sequence positions in which two or more alternatives were observed in at least 20% of genome collection. The number of intra-species SNPs obtained for each *Lactobacillus* species was converted into SNPs per Mbp to account for variation in genome length.

In a similar fashion, both concatenated and individual gene nucleotide sequences comprised within the *L. crispatus* core genome were aligned with Mafft v7.453 and parsed with the SNP-site software. For these analyses, the genome sequence of the most divergent *L. crispatus* strain (assembly number GCF_015669875.1) was used as reference sequence. Synonymous and non-synonymous nucleotide variations were discriminated using the ParaAT 2.0 software (49) combined with KaKs Calculator 3.0 toolkit (50). Whole-genome SNPs were extracted by combining the short-reads aligner BWA and the VarScan tool (version 2.3.6)

Carbohydrate growth assay

In vitro growth assays with different carbon sources, such as starch, amylopectin, pullulan, maltodextrin, glycogen, and mucin, were performed on selected *L. crispatus* strains, i.e., LB97, LMG11440, LMG18200, and LMG11415. In detail, the four *L. crispatus* strains were cultivated overnight on a semisynthetic MRS medium supplemented with 0.05% (w/vol) L-cysteine hydrochloride at 37 °C under anaerobic conditions. Subsequently, cells were diluted in MRS without glucose to obtain an $OD_{600\ nm}$=1, and 15 µl of the diluted cells were inoculated in 135 µl of MRS without glucose supplemented with 1% (wt/vol) of a particular sugar in a 96-well microtiter plate and incubated in an anaerobic cabinet. Specifically, each carbohydrate was dissolved in MRS without glucose previously sterilized by autoclaving at 121°C for 15 min. Subsequently, the obtained solutions were sterilized using a 0.2 µm filter size before use. Cell growth was evaluated by monitoring the optical density at 600 nm using a plate reader (Biotek, VT, USA). Each plate was read in discontinuous mode, with absorbance readings performed thrice at 3-min intervals after 48h of growth, and each reading was ahead of 30s of shaking at medium speed. Cultures were performed in triplicates for each strain, and the resulting growth data were expressed as the average OD_{600nm} of these independent biological replicates. Carbohydrates

tested in this study were purchased from Merck (Germany) and Fisher Scientific, ACROS Organics (USA) and include soluble starch from potato, amylopectin from maize, pullulan, maltodextrin, glycogen from beef liver, mucin from porcine stomach. The semisynthetic MRS medium was used as the control medium.

Co-culture using a bioreactor system

The four selected *L. crispatus* strains (reported above) were grown anaerobically at 37°C for 24h in simulated vaginal fluid (SVF) (24) to adapt the microorganisms to the medium. Next, revitalized cells were inoculated in a bioreactor system (Solaris Biotech Solutions, Italy) to obtain a final concentration per bacterial strain of $5x10^6$ cells/mL in 400 mL of SVF. The co-culture of the four *L. crispatus* strains was performed with a non-continuous supply of the growth medium for the first 12h to stabilize the microbial community. Subsequently, the cultivation was shifted to a continuous mode to provide fresh SVF medium and continued for 48h under anaerobic conditions at 37°C with a mechanical agitation set at 180 rpm. In addition, the pH was maintained at 4.5 by adding 2.5M NaOH to mimic the pH of the human vaginal environment (24). During bacterial growth, culture aliquots were collected at 10h, 24h, and 48h.

DNA extraction and shotgun metagenomic sequencing

Each aliquot collected from bioreactor cultivation was subjected to DNA extraction using the ZymoBIOMICS DNA Miniprep Kit (Zymo Research, D4300) following the manufacturer's instructions. Then, after assessing DNA concentration and purity using a BioPhotometer D30 (Eppendorf, Germany), each DNA sample was sequenced by GenProbio srl, Parma, Italy (www.genprobio.com) employing next-generation sequencing technique (shotgun metagenomic sequencing). The preparation of DNA libraries was performed using the Nextera XT DNA sample preparation kit (Illumina, San Diego, CA) according to the manufacturer's

instructions, using 1 ng of DNA from each metagenomic sample. The isolated DNA underwent fragmentation, adapter ligation, and purification. The ready-to-go libraries were pooled equimolarly and diluted to a sequencing concentration of 650 pM. On-board DNA denaturation and sequencing were performed on a NextSeq 2000 instrument (Illumina, San Diego, CA), according to the manufacturer's instructions, using the 2x150 bp NextSeq 1000/2000 P2 Reagents (300 Cycles) v3 and spike-in of 1 % PhiX control library. Whole-metagenome shotgun (WMGS) sequencing of the three bioreactor culture aliquots produced an average of 21,861,838 ± 7,995,330 paired-end 150 bp reads per sample. Raw metagenomic sequencing reads were trimmed and quality filtered with fastq-mcf software supplied by Illumina Inc (minimum mean quality score, 20; window size, 5 bp; and minimum length, 100 bp). Following quality filtering, an average of 18,662,746 ± 5,671,267 quality-filtered microbial reads per sample were retained (Supplementary Table S8).

L. crispatus strain-level profiling of the bioreactor-derived cultures

To disentangle the different *L. crispatus* strains in the co-culture aliquots collected at different time points (10h, 24h, and 48h) during bioreactor growth, the filtered metagenomic reads obtained from each shotgun sequencing effort were mapped against specific distinctive regions of every *L. crispatus* genome using the software BBMap (https://sourceforge.net/projects/bbmap/) with 100% homology (perfectmode=t flag). Notably, to avoid mis-mapping of the metagenomic short reads, it was used *L. crispatus* genomes that returned pairwise ANI values < 98%, as advised in a previous qualified study (https://drep.readthedocs.io/en/latest/choosing_parameters.html) (51). In detail, to identify suitable discriminative genes, the whole set of genes unique to each *L. crispatus* strain detected in the PGAP analysis were mapped against the combined genomes of all strains using the Bowtie2 --very-sensitive mode (52). Genes that did

not return any hits other than those corresponding to the genome to which they belong were retained as candidate strain-specific marker genes. This selection was then manually inspected to exclude genes corresponding to transposases, phage genes, and genes located alongside the contig ends. This procedure identified a set of roughly ten unique marker genes for each *L. crispatus* strain that were used in downstream analyses on the bioreactor-derived metagenomic reads (Supplementary Table S7). A proxy measure of each strain abundance was calculated by normalizing the mapped read count on the corresponding marker gene length and library size using the RPKM mathematical formula [(10^9 * Number of mapped reads to a gene)/(Total mapped reads * gene length in base-pairs]). Moreover, the set of genes associated uniquely with each of the four co-cultivated strains was functionally investigated to discover potential accessory protein-encoding sequences conferring peculiar growth abilities in the cultivation medium. To this scope, we employed the MetaCyc database (https://metacyc.org/), which allowed us to assign a detailed functional annotation to each scrutinized gene. In addition, the Transporter Classification Database (TCDB) was exploited to characterize transport systems and identify their possible substrates (https://tcdb.org/).

Statistical analysis

The software SPSS version 25 and OriginPro version 2023 (www.ibm.com/software/it/analytics/spss/) (https://www.originlab.com/) were used for statistical data analyses and graphing. One-way ANOVA with Bonferroni correction was used to determine the statistical significance of differences in the OD_{600} measures (growth assay) and normalized read counts (bioreactor-based co-cultivation experiment). AlphaFold (53) and PyMOL software (https://pymol.org/2/) were used to observe SNPs within the predicted 3D protein structure of the pullulanase type II gene derived from the *L. crispatus* LB97.

Data availability

Genome sequences of the 41 newly sequenced *L. crispatus* were deposited in NCBI-SRA (Short Read Archive) repository with accession number PRJNA947599.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

We thank GenProbio Srl for the financial support to the Laboratory of Probiogenomics. The cost of the equipment (Sirius One MALDI-TOF Bioanalyzer) used for this experimental investigation was partly supported by the University of Parma through the Scientific Instrumentation Upgrade Programme 2020. Part of this research is conducted using the High-Performance Computing (HPC) facility of the University of Parma.

Funding

C.A. is supported by Fondazione Cariparma, Parma, Italy.

References

1. Blum HE. The human microbiome. Adv Med Sci [Internet]. 2017 Sep 1 [cited 2023 Mar 16];62(2):414–20. Available from: https://pubmed.ncbi.nlm.nih.gov/28711782/

2. Warner BB. The contribution of the gut microbiome to neurodevelopment and neuropsychiatric disorders. Pediatr Res. 2019 Jan 1;85(2):216–24.

3. Bifidobacteria: insights into the biology of a key microbial group of early life gut microbiota - Search [Internet]. [cited 2023 Feb 3]. Available from: https://www.bing.com/search?q=Bifidobacteria%3A+insights+into+the+biology+of+a+key+microbial+group+of+early+life+gut+microbiota&cvid=b6d628128deb4e00a9ea38a06b32a4b0&aqs=edge..69i57j69i58.294j0j9&FORM=ANAB01&PC=U531

4. Turroni F, Rizzo SM, Ventura M, Bernasconi S. Cross-talk between the infant/maternal gut microbiota and the endocrine system: a promising topic of research. Microbiome Research Reports [Internet]. 2022 Mar 31 [cited 2023 Feb 3];1(2):14. Available from: https://www.oaepublish.com/mrr/article/view/4754

5. Sommer F, Bäckhed F. The gut microbiota — masters of host development and physiology. Nature Reviews Microbiology 2013 11:4 [Internet]. 2013 Feb 25 [cited 2023 Feb 3];11(4):227–38. Available from: https://www.nature.com/articles/nrmicro2974

6. Tarracchini C, Fontana F, Mancabelli L, Lugli GA, Alessandri G, Turroni F, et al. Gut microbe metabolism of small molecules supports human development across the early stages of life. Front Microbiol [Internet]. 2022 Sep 13 [cited 2023 Feb 3];13. Available from: https://pubmed.ncbi.nlm.nih.gov/36177457/

7. Strati F, Facciotti F. Gut microbiota-derived metabolites in host physiology. Metabolomics Perspectives: From Theory to Practical Application [Internet]. 2022 Jan 1 [cited 2023 Feb 3];515–34. Available from: https://www.researchgate.net/publication/367747520_Gut_microbiota-derived_metabolites_in_host_physiology

8. Tachedjian G, Aldunate M, Bradshaw CS, Cone RA. The role of lactic acid production by probiotic Lactobacillus species in vaginal health. Res Microbiol [Internet]. 2017 Nov 1 [cited 2023 Mar 6];168(9–10):782–92. Available from: https://pubmed.ncbi.nlm.nih.gov/28435139/

9. Lepargneur JP. Lactobacillus crispatus as biomarker of the healthy vaginal tract. Ann Biol Clin (Paris) [Internet]. 2016 Jul 1 [cited 2023 Mar 6];74(4):421–7. Available from: https://pubmed.ncbi.nlm.nih.gov/27492695/

10. Sanozky-Dawes R, Barrangou R. Lactobacillus , glycans and drivers of health in the vaginal microbiome. Microbiome Research Reports [Internet]. 2022 May 13 [cited 2023 Feb 3];1(3):18. Available from: https://www.oaepublish.com/mrr/article/view/4861

11. Stapleton AE, Au-Yeung M, Hooton TM, Fredricks DN, Roberts PL, Czaja CA, et al. Randomized, placebo-controlled phase 2 trial of a Lactobacillus crispatus probiotic given intravaginally for prevention of recurrent urinary tract infection. Clin Infect Dis [Internet]. 2011 May 15 [cited 2023 Feb 14];52(10):1212–7. Available from: https://pubmed.ncbi.nlm.nih.gov/21498386/

12. Cohen CR, Wierzbicki MR, French AL, Morris S, Newmann S, Reno H, et al. Randomized Trial of Lactin-V to Prevent Recurrence of Bacterial Vaginosis. N Engl J Med [Internet]. 2020 May 14 [cited 2023 Mar 6];382(20):1906–15. Available from: https://pubmed.ncbi.nlm.nih.gov/32402161/

13. Bohbot JM, Daraï E, Bretelle F, Brami G, Daniel C, Cardot JM. Efficacy and safety of vaginally administered lyophilized Lactobacillus crispatus IP 174178 in the prevention of bacterial vaginosis recurrence. J Gynecol Obstet Hum Reprod [Internet]. 2018 Feb 1 [cited 2023 Mar 6];47(2):81–6. Available from: https://pubmed.ncbi.nlm.nih.gov/29196153/

14. Mändar R, Sõerunurk G, Štšepetova J, Smidt I, Rööp T, Kõljalg S, et al. Impact of Lactobacillus crispatus-containing oral and vaginal probiotics on vaginal health: a randomised double-blind placebo controlled clinical trial. Benef Microbes [Internet]. 2023 Mar 1 [cited 2023 Mar 6];1–10. Available from: https://pubmed.ncbi.nlm.nih.gov/36856121/

15. Puebla-Barragan S, Watson E, van der Veer C, Chmiel JA, Carr C, Burton JP, et al. Interstrain Variability of Human Vaginal Lactobacillus crispatus for Metabolism of Biogenic Amines and Antimicrobial Activity against Urogenital Pathogens. Molecules [Internet]. 2021 Aug 1 [cited 2023 Feb 14];26(15). Available from: https://pubmed.ncbi.nlm.nih.gov/34361691/

16. Argentini C, Fontana F, Alessandri G, Lugli GA, Mancabelli L, Ossiprandi MC, et al. Evaluation of Modulatory Activities of Lactobacillus crispatus Strains in the Context of the Vaginal Microbiota. Microbiol Spectr [Internet]. 2022 Apr 27 [cited 2023 Feb 3];10(2). Available from: https://pubmed.ncbi.nlm.nih.gov/35266820/

17. van der Veer C, Hertzberger RY, Bruisten SM, Tytgat HLP, Swanenburg J, de Kat Angelino-Bart A, et al. Comparative genomics of human Lactobacillus crispatus isolates reveals genes for glycosylation and glycogen degradation: Implications for in vivo dominance of the vaginal microbiota. Microbiome [Internet]. 2019 Mar 29 [cited 2023 Feb 14];7(1):1–14. Available from: https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0667-9

18. Ojala T, Kankainen M, Castro J, Cerca N, Edelman S, Westerlund-Wikström B, et al. Comparative genomics of Lactobacillus crispatus suggests novel mechanisms for the competitive exclusion of Gardnerella vaginalis. BMC Genomics [Internet]. 2014 Dec 5 [cited 2023 Feb 14];15(1). Available from: /pmc/articles/PMC4300991/

19. Abdelmaksoud AA, Koparde VN, Sheth NU, Serrano MG, Glascock AL, Fettweis JM, et al. Comparison of Lactobacillus crispatus isolates from Lactobacillus-dominated vaginal microbiomes with isolates from microbiomes containing bacterial vaginosis-associated bacteria. Microbiology (Reading) [Internet]. 2016 Feb 1 [cited 2023 May 11];162(3):466–75. Available from: https://pubmed.ncbi.nlm.nih.gov/26747455/

20. Mendes-Soares H, Suzuki H, Hickey RJ, Forneya LJ. Comparative functional genomics of Lactobacillus spp. reveals possible mechanisms for specialization of vaginal lactobacilli to their environment. J Bacteriol [Internet]. 2014 [cited 2023 May 11];196(7):1458–70. Available from: https://pubmed.ncbi.nlm.nih.gov/24488312/

21. Zhang Q, Zhang L, Ross P, Zhao J, Zhang H, Chen W. Comparative Genomics of Lactobacillus crispatus from the Gut and Vagina Reveals Genetic Diversity and Lifestyle Adaptation. Genes (Basel) [Internet]. 2020 Apr 1 [cited 2023 Feb 23];11(4). Available from: /pmc/articles/PMC7230607/

22. Mancabelli L, Mancino W, Lugli GA, Milani C, Viappiani A, Anzalone R, et al. Comparative genome analyses of Lactobacillus crispatus isolated from different ecological niches reveal an environmental adaptation of this species to the human vaginal environment. Appl Environ Microbiol [Internet]. 2021 Apr 1 [cited 2023 Feb 23];87(8):1–21. Available from: https://pubmed.ncbi.nlm.nih.gov/33579685/

23. France MT, Mendes-Soares H, Forney LJ. Genomic Comparisons of Lactobacillus crispatus and Lactobacillus iners Reveal Potential Ecological Drivers of Community Composition in the Vagina. Appl Environ Microbiol [Internet]. 2016 Dec 12 [cited 2023 Feb 23];82(24):7063. Available from: /pmc/articles/PMC5118917/

24. Pan M, Hidalgo-Cantabrana C, Goh YJ, Sanozky-Dawes R, Barrangou R. Comparative Analysis of Lactobacillus gasseri and Lactobacillus crispatus Isolated From Human Urogenital and Gastrointestinal Tracts. Front Microbiol. 2020 Jan 22;10:3146.

25. Rousset F, Cabezas-Caballero J, Piastra-Facon F, Fernández-Rodríguez J, Clermont O, Denamur E, et al. The impact of genetic diversity on gene essentiality within the Escherichia coli species. Nature Microbiology 2021 6:3 [Internet]. 2021 Jan 18 [cited 2023 Mar 6];6(3):301–12. Available from: https://www.nature.com/articles/s41564-020-00839-y

26. Juhas M, Eberl L, Glass JI. Essence of life: essential genes of minimal genomes. Trends Cell Biol [Internet]. 2011 Oct [cited 2023 Mar 6];21(10):562–8. Available from: https://pubmed.ncbi.nlm.nih.gov/21889892/

27. Martínez-Carranza E, Barajas H, Alcaraz LD, Servín-González L, Ponce-Soto GY, Soberón-Chávez G. Variability of bacterial essential genes among closely related bacteria: The case of Escherichia coli. Front Microbiol. 2018 May 29;9(MAY):1059.

28. 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." Proc Natl Acad Sci U S A. 2005 Sep 27;102(39):13950–5.

29. Muthukumarasamy U, Preusse M, Kordes A, Koska M, Schniederjans M, Khaledi A, et al. Single-Nucleotide Polymorphism-Based Genetic Diversity Analysis of Clinical Pseudomonas aeruginosa Isolates. Genome Biol Evol [Internet]. 2020 [cited 2023 May 15];12(4):396–406. Available from: https://pubmed.ncbi.nlm.nih.gov/32196089/

30. Shakya M, Ahmed SA, Davenport KW, Flynn MC, Lo CC, Chain PSG. Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. Scientific Reports 2020 10:1 [Internet]. 2020 Feb 3 [cited 2023 May 17];10(1):1–15. Available from: https://www.nature.com/articles/s41598-020-58356-1

31. Tester R, Al-Ghazzewi FH. Intrinsic and extrinsic carbohydrates in the vagina: A short review on vaginal glycogen. Int J Biol Macromol. 2018 Jun 1;112:203–6.

32. Zhang J, Li L, Zhang T, Zhong J. Characterization of a novel type of glycogen-degrading amylopullulanase from Lactobacillus crispatus. Appl Microbiol Biotechnol [Internet]. 2022 Jun 1 [cited 2023 Mar 2];106(11):4053–64. Available from: https://pubmed.ncbi.nlm.nih.gov/35612627/

33. Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, et al. The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Res [Internet]. 2018 Jan 1 [cited 2021 Sep 21];46(D1):D633–9. Available from: https://pubmed.ncbi.nlm.nih.gov/29059334/

34. Linares D, Michaud P, Delort AM, Traïkia M, Warrand J. Catabolism of L-ascorbate by Lactobacillus rhamnosus GG. J Agric Food Chem [Internet]. 2011 Apr 27 [cited 2023 Mar 17];59(8):4140–7. Available from: https://pubmed.ncbi.nlm.nih.gov/21401096/

35. Mirmonsef P, Hotton AL, Gilbert D, Burgad D, Landay A, Weber KM, et al. Free glycogen in vaginal fluids is associated with Lactobacillus colonization and low vaginal pH. PLoS One. 2014 Jul 17;9(7).

36. Werner G, Fleige C, Feßler AT, Timke M, Kostrzewa M, Zischka M, et al. Improved identification including MALDI-TOF mass spectrometry analysis of group D streptococci from bovine mastitis and subsequent molecular characterization of corresponding Enterococcus faecalis and Enterococcus faecium isolates. Vet Microbiol [Internet]. 2012 Nov 9 [cited 2023 Feb 3];160(1–2):162–9. Available from: https://pubmed.ncbi.nlm.nih.gov/22677481/

37. Schulthess B, Bloemberg G v., Zbinden R, Böttger EC, Hombach M. Evaluation of the Bruker MALDI Biotyper for identification of Gram-positive rods: development of a diagnostic algorithm for the clinical laboratory. J Clin Microbiol [Internet]. 2014 [cited 2023 Feb 3];52(4):1089–97. Available from: https://pubmed.ncbi.nlm.nih.gov/24452159/

38. Lugli GA, Milani C, Mancabelli L, Van Sinderen D, Ventura M. MEGAnnotator: a user-friendly pipeline for microbial genomes assembly and annotation. FEMS Microbiol Lett [Internet]. 2016 Apr 1 [cited 2023 Jan 30];363(7):49. Available from: https://academic.oup.com/femsle/article/363/7/fnw049/2197823

39. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology [Internet]. 2012 May 1 [cited 2021 Nov 11];19(5):455–77. Available from: https://pubmed.ncbi.nlm.nih.gov/22506599/

40. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics [Internet]. 2010 Mar 8 [cited 2021 Dec 9];11. Available from: https://pubmed.ncbi.nlm.nih.gov/20211023/

41. Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics [Internet]. 2012 Jan [cited 2021 Dec 9];28(1):125–6. Available from: https://pubmed.ncbi.nlm.nih.gov/22039206/

42. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res [Internet]. 1997 Mar 1 [cited 2021 Dec 9];25(5):955–64. Available from: https://pubmed.ncbi.nlm.nih.gov/9023104/

43. Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res [Internet]. 2007 May [cited 2021 Dec 9];35(9):3100–8. Available from: https://pubmed.ncbi.nlm.nih.gov/17452365/

44. Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J. PGAP: pan-genomes analysis pipeline. Bioinformatics [Internet]. 2012 Feb [cited 2021 Dec 9];28(3):416–8. Available from: https://pubmed.ncbi.nlm.nih.gov/22130594/

45. Tarracchini C, Lugli GA, Mancabelli L, Milani C, Turroni F, Ventura M. Assessing the Genomic Variability of Gardnerella vaginalis through Comparative Genomic Analyses: Evolutionary and Ecological Implications. Appl Environ Microbiol [Internet]. 2020 Jan 1 [cited 2021 Nov 11];87(1):1–16. Available from: https://pubmed.ncbi.nlm.nih.gov/33097505/

46. Lugli GA, Duranti S, Albert K, Mancabelli L, Napoli S, Viappiani A, et al. Unveiling Genomic Diversity among Members of the Species Bifidobacterium pseudolongum, a Widely Distributed Gut Commensal of the Animal Kingdom. Appl Environ Microbiol [Internet]. 2019 Apr 1 [cited 2023 Feb 7];85(8). Available from: https://pubmed.ncbi.nlm.nih.gov/30737347/

47. Katoh K, Misawa K, Kuma KI, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res [Internet]. 2002 Jul 15 [cited 2021 Dec 9];30(14):3059–66. Available from: https://pubmed.ncbi.nlm.nih.gov/12136088/

48. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom [Internet]. 2016 Apr 1 [cited 2023 Feb 7];2(4):e000056. Available from: https://pubmed.ncbi.nlm.nih.gov/28348851/

49. Zhang Z, Xiao J, Wu J, Zhang H, Liu G, Wang X, et al. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. Biochem Biophys Res Commun [Internet]. 2012 Mar 23 [cited 2023 Mar 15];419(4):779–81. Available from: https://pubmed.ncbi.nlm.nih.gov/22390928/

50. Zhang Z. KaKs_Calculator 3.0: Calculating Selective Pressure on Coding and Non-coding Sequences. Genomics Proteomics Bioinformatics. 2022 Jun 1;20(3):536–40.

51. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nature Biotechnology 2021 39:6 [Internet]. 2021 Jan 18 [cited 2022 Sep 26];39(6):727–36. Available from: https://www.nature.com/articles/s41587-020-00797-0

52. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods [Internet]. 2012 Apr [cited 2021 Nov 11];9(4):357–9. Available from: https://pubmed.ncbi.nlm.nih.gov/22388286/

53. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021 596:7873 [Internet]. 2021 Jul 15 [cited 2023 May 17];596(7873):583–9. Available from: https://www.nature.com/articles/s41586-021-03819-2

Chapter 6

Assessing the Genomic Variability of *Gardnerella vaginalis* through Comparative Genomic Analyses: Evolutionary and Ecological Implications

Chiara Tarracchini, Gabriele Andrea Lugli, Leonardo Mancabelli, Christian Milani, Francesca Turroni, Marco Ventura

The results of this chapter were published in Applied and Environmental Microbiology, 2020 Dec;
doi: 10.1128/AEM.02188-20.

Abstract

Gardnerella vaginalis is described as a common anaerobic vaginal bacterium whose presence may correlate with vaginal dysbiotic conditions. In the current study, we performed phylogenomic analyses of 72 *G. vaginalis* genome sequences, revealing noteworthy genome differences underlying a polyphyletic organization of this taxon. Particularly, the genomic survey revealed that this species may actually include nine distinct genotypes (GGtype1 to GGtype9). Furthermore, the observed link between sialidase and phylogenomic grouping provided clues of a connection between virulence potential and the evolutionary history of this microbial taxon. Specifically, based on the outcomes of these in silico analyses, GGtype3, GGtype7, GGtype8, and GGtype9 appear to have virulence potential since they exhibited the sialidase gene in their genomes. Notably, the analysis of 34 publicly available vaginal metagenomic samples allowed us to trace the distribution of the nine *G. vaginalis* genotypes identified in this study among the human population, highlighting how differences in genetic makeup could be related to specific ecological properties. Furthermore, comparative genomic analyses provided details about the *G. vaginalis* pan- and core genome contents, including putative genetic elements involved in the adaptation to the ecological niche as well as many putative virulence factors. Among these putative virulence factors, particularly noteworthy genes identified were the gene encoding cholesterol-dependent cytolysin (CDC) toxin vaginolysin and genes related to microbial biofilm formation, iron uptake, adhesion to the vaginal epithelium, as well as macrolide antibiotic resistance.

IMPORTANCE. The identification of nine different genotypes among members of *G. vaginalis* allowed us to distinguish an uneven distribution of virulence-associated genetic traits within this taxon and thus suggest the potential occurrence of putative pathogen and commensal *G. vaginalis* strains. These findings, coupled

with metagenomics microbial profiling of human vaginal microbiota, permitted us to get insights into the distribution of the genotypes among the human population, highlighting the presence of different structural communities in terms of *G. vaginalis* genotypes.

For Supplementary Materials see the article published in Applied and Environmental Microbiology

INTRODUCTION

The human female reproductive tract harbors trillions of bacteria that play an important role in the health of women (1). In particular, human vaginal microbiota are believed to exert a preventive action against several diseases, such as bacterial vaginosis (BV), sexually transmitted diseases (STDs), and urinary tract infections (2–5). In this context, members of the *Lactobacillus* genus are generally dominant in the vaginal microenvironment of healthy women and exploit their beneficial role(s) through lactic acid production that keeps low pH and provide protection to the host against pathogenic bacteria (6–8).

In recent years, the composition of women's vaginal microbiota has been investigated by means of next-generation DNA sequencing techniques, revealing that vaginal bacterial communities, i.e., vaginal microbiota, can be classified from three to nine ecotypes according to their specific microbial composition (9). In this context, it has been proposed that the vaginal microbiota of asymptomatic women from four ethnic groups could be clustered into five community-state types (CSTs) (9). Notably, CST I, CST II, CST III, and CST V were dominated by various species of *Lactobacillus*, i.e., *Lactobacillus crispatus*, *Lactobacillus gasseri*, *Lactobacillus iners*, and *Lactobacillus jensenii*, respectively, while CST IV was mainly constituted by obligated anaerobic bacteria, also including members of the *Prevotella*, *Atopobium*, and *Gardnerella* genera. Among the latter genus, a single species has been so far described, i.e., Gardnerella vaginalis, represented by Gram-positive, anaerobic, non-spore-forming bacteria commonly identified in the vaginal environment (10).

Great interest revolves around *G. vaginalis* since this microorganism was frequently detected as a dominant microorganism in chronic and acute BV incidence (11), which is an aberrant condition characterized by a shift of the vaginal microbiota composition (*Lactobacillus* dominated) toward a more diversified microbial

community (12). It has been demonstrated that *G. vaginalis* cells possess the ability to adhere to the vaginal epithelium and develop a characteristic microbial biofilm (13, 14), enabling it to colonize the vaginal tract efficiently. In addition, *G. vaginalis* can produce other virulence factors, such as sialidase, which has been strongly linked with microbial biofilm production (15, 16), and cholesterol-dependent cytolysin (CDC) family toxin vaginolysin (17). However, it has also been observed that presence of *G. vaginalis* in the vaginal microbiota does not always imply BV (18). For this reason, several efforts were made to highlight which genomic differences could discriminate pathogenic from commensal strains (11, 19, 20). Nevertheless, the role of *G. vaginalis* in the pathogenesis of BV is still far from being fully understood.

Since its discovery in 1955, *G. vaginalis* was named *Haemophilus vaginalis* (10) and later, it was designated *Corynebacterium vaginale* (21). Afterward, taxonomic studies confirmed the need to introduce the new *Gardnerella* genus, also showing its taxonomic relatedness to the *Bifidobacterium* genus (22,23). To date, *G. vaginalis* is taxonomically placed within the *Bifidobacteriaceae* family, and it is considered the only species of the *Gardnerella* genus. However, several studies have reported the existence of genetic heterogeneity among the various members of this genus (24-26). Here, we carried out an exhaustive comparative genome analysis based on 72 publicly available genomic sequences of *G. vaginalis*, aiming to investigate the genomic variability of this taxon. Moreover, phylogenomics analyses were carried out to highlight the phylogenetic relationships of *G. vaginalis* with the other members of the *Bifidobacteriaceae* family. Finally, the screening of 34 publicly available vaginal shotgun metagenomic data sets allowed us to investigate the distribution of the here-identified *G. vaginalis* genotypes among the human population.

RESULTS AND DISCUSSION

General genome features of *Gardnerella vaginalis*

In order to perform an exhaustive comparative genomic analysis of the *G. vaginalis* species, all the publicly available genome sequences of this taxon were retrieved from the NCBI database (Table 1). Notably, chromosomes of *G. vaginalis* used in this work were carefully selected, resulting in one of the largest high-quality databases developed to date, encompassing 72 *G. vaginalis* genomes (see Materials and Methods). The predicted average genomic GC content was 41.8%, a lower value than the other members of the *Bifidobacteriaceae* family (60.2% for the bifidobacterial strains and 52.9% for other genera of the bifidobacterial strains and 52.9% for other genera of the *Bifidobacteriaceae* family) (27). Interestingly, the GC content showed low variability among analyzed strains, except for the CMW7778B chromosome, which deviates from the genomes of the other strains, with a GC content of 38%. As shown by previous studies, these findings allowed researchers to suggest that the adaptation to a limited niche complexity, along with a constant environmental temperature, may have affected the GC content of *G. vaginalis* genomes (28). In fact, *G. vaginalis* strains have been isolated so far only from the human urogenital tract, thus showing a restricted ecological niche whose temperature is maintained to be almost constant. In contrast, members of the *Bifidobacterium* genus that colonize a wide variety of ecological niches, including the gut of homeothermic and heterothermic animals, exhibited a higher GC content level (27).

Table 1. General genome features of *G. vaginalis* **[a]NA, Not Available**

GARDNERELLA VAGINALIS STRAIN	ENA assembly no.	Genome status	Genome size (Mb)	GC content (%)	No. of CDS	No. of rRNA loci	No. of tRNA genes	Virulence gene(s)	Isolation source
5-1	GCA_000176495.1	Draft	1.6728	42.0	1,273	1	45	*vly*	Vagina
41V	GCA_000165635.2	Draft	1.6594	41.3	1,277	1	45	*vly*	Vagina
PSS_7772B	GCA_001546485.1	Draft	1.5967	42.9	1,169	1	44	*vly*	Urine
KA00225	GCA_002896555.1	Draft	1.6700	40.8	1,187	2	45	*vly*	Vagina
101	GCA_000165615.2	Draft	1.5275	43.4	1,163	2	45	*vly*	NA[a]
CMW7778B	GCA_001563665.1	Draft	1.6026	38.0	1,150	1	44	*vly*	Vagina
N165	GCA_003408785.1	Draft	1.7116	41.4	1,344	2	44	*vly*, *sld*	Vaginal mucus
1400E	GCA_000263495.1	Draft	1.7163	41.2	1,331	3	44	*vly*	Vagina
1500E	GCA_000263595.1	Draft	1.5482	43.0	1,157	3	45	*vly*	Vagina
55152	GCA_000263475.1	Draft	1.6432	41.3	1,244	1	45	*vly*	Vagina
GED7760B	GCA_001546455.1	Draft	1.4892	43.3	1,123	1	45	*sld*	Vagina
UGENT 09.07	GCA_003397665.1	Draft	1.7238	41.1	1,293	1	47	*vly*	Vagina
00703C2MASH	GCA_000263515.1	Draft	1.5467	42.3	1,185	2	45	*vly*	Vagina
49145	GCA_003034925.1	Draft	1.7014	41.2	1,325	1	45	*vly*, *sld*	Vagina
ATCC 49145	GCA_001913835.1	Draft	1.7069	41.2	1,361	2	45	*vly*, *sld*	Vagina
GED7275B	GCA_001546445.1	Draft	1.5079	42.5	1,139	1	45	*vly*	Vagina
UMB0061	GCA_002861165.1	Draft	1.7422	41.2	1,387	2	45	*vly*, *sld*	Catheter
0288E	GCA_000263555.1	Draft	1.7088	41.2	1,338	1	45	*vly*	Vagina
284V	GCA_000263435.1	Draft	1.6508	41.2	1,280	2	45	*vly*	Vagina
00703BMASH	GCA_000263615.1	Draft	1.5661	42.3	1,227	1	45	*vly*, *sld*	Vagina
JCM 11026	GCA_004336685.1	Draft	1.6571	41.3	1,225	1	45	*vly*, *sld*	Vagina
6420B	GCA_000263575.1	Draft	1.4936	42.2	1,122	1	45	*vly*	Vagina
UMB0032B	GCA_002862005.1	Draft	1.7451	41.2	1,382	1	45	*vly*, *sld*	Catheter
315-A	GCA_000214315.2	Draft	1.6533	41.4	1,298	1	45	*vly*, *sld*	Vaginal
NR010	GCA_003408845.1	Draft	1.6227	45.5	1,181	3	45	*vly*, *sld*	Vaginal mucus
UMB0833	GCA_002861885.1	Draft	1.6203	42.1	1,273	3	45	*sld*	Catheter
6119V5	GCA_000263655.1	Draft	1.4996	43.3	1,117	1	45	*vly*	Vagina
3549624	GCA_001049785.1	Draft	1.7323	41.4	1,298	2	45	*vly*, *sld*	Vagina
00703DMASH	GCA_000263635.1	Draft	1.4908	43.4	1,121	1	45	*vly*	Vagina
14018C	GCA_004336715.1	Draft	1.6578	41.3	1,232	1	45	*vly*, *sld*	NA[a]
UMB0032A	GCA_002862015.1	Draft	1.7455	41.2	1,383	2	45	*vly*, *sld*	Catheter
UMB0770	GCA_002861945.1	Draft	1.6960	41.2	1,323	1	45	*vly*, *sld*	Catheter
UMB0775	GCA_002861925.1	Draft	1.7436	41.2	1,397	2	45	*vly*, *sld*	Catheter
GS 9838-1	GCA_003397705.1	Draft	1.6221	41.9	1,231	2	45	*vly*	Vagina
UMB1686	GCA_002884775.1	Draft	1.5106	43.3	1,124	2	45	*vly*, *sld*	Catheter
DNF01149	GCA_002894105.1	Draft	1.7247	41.2	1,362	3	45	*vly*, *sld*	Vagina
N101	GCA_003369895.1	Draft	1.5430	42.4	1,205	1	45	*vly*, *sld*	Vaginal swab

UMB0233	GCA_002862045.1	Draft	1.6424	41.2	1,292	1	45	*vly*, *sld*	Catheter
14019_METR	GCA_001278345.1	Draft	1.6611	41.3	1,317	1	45	*vly*, *sld*	NA[a]
W11	GCA_003369875.1	Draft	1.5667	42.3	1,213	1	45	*sld*	Vaginal swab
N95	GCA_003369965.1	Draft	1.5225	42.4	1,183	2	45	*vly*, *sld*	Vaginal swab
UMB0682	GCA_002862065.1	Draft	1.6013	42.1	1,234	1	45	*vly*	Catheter
N153	GCA_003369935.1	Draft	1.5418	42.4	1,167	1	45	*vly*, *sld*	Vaginal swab
N72	GCA_003408815.1	Draft	1.6429	41.9	1,249	1	45	*vly*	Vaginal mucus
N160	GCA_003408775.1	Draft	1.5097	43.3	1,119	3	43	*vly*, *sld*	Vaginal mucus
UGENT 18.01	GCA_003397585.1	Draft	1.5143	42.5	1,144	1	45	*sld*	Vagina
UMB0768	GCA_002884835.1	Draft	1.6748	41.3	1,319	2	45	*vly*, *sld*	Catheter
UMB1642	GCA_002884795.1	Draft	1.6288	41.8	1,233	2	45	*vly*	Catheter
UMB0264	GCA_002884875.1	Draft	1.5151	42.3	1,155	2	45	*vly*	Catheter
UMB0913	GCA_002861145.1	Draft	1.5136	42.1	1,140	1	45	*vly*	Catheter
UMB0170	GCA_002884855.1	Draft	1.5147	42.3	1,153	2	45	*vly*	Catheter
UMB0912	GCA_002861125.1	Draft	1.5138	42.1	1,140	1	45	*vly*	Catheter
UMB0830	GCA_002861905.1	Draft	1.5592	42.3	1,224	1	45	*vly*, *sld*	Catheter
75712	GCA_000263535.1	Draft	1.6730	41.3	1,302	2	45	*vly*	Vagina
UMB0386	GCA_002861965.1	Draft	1.6757	41.2	1,323	2	45	*vly*, *sld*	Catheter
UGENT 25.49	GCA_003397605.1	Draft	1.6586	41.2	1,246	2	45	*vly*, *sld*	Vagina
GS 10234	GCA_003397745.1	Draft	1.5890	41.9	1,181	2	45	*vly*	Vagina
UGENT 09.48	GCA_003397635.1	Draft	1.4709	42.2	1,095	1	45	*vly*	Vagina
UGENT 21.28	GCA_003397615.1	Draft	1.5479	42.5	1,189	1	45	*sld*	Vagina
UMB0298	GCA_002861975.1	Draft	1.6760	41.2	1,319	2	45	*vly*, *sld*	Catheter
ATCC 14018	GCA_003397685.1	Draft	1.6620	41.3	1,248	1	45	*vly*, *sld*	Vagina
FDAARGOS_296	GCA_002206225.2	Draft	1.7710	41.3	1,375	2	45	*vly*, *sld*	NA[a]
N144	GCA_003408835.1	Draft	1.5824	42.3	1,217	1	45	*vly*, *sld*	Vaginal mucus
GH015	GCA_003408745.1	Draft	1.5756	41.0	1,173	1	43	*vly*, *sld*	Vaginal mucus
JCM 11026	GCA_001042655.1	Complete	1.6674	41.3	1,244	2	45	*vly*, *sld*	Vagina
NCTC10287	GCA_900637625.1	Complete	1.6674	41.4	1,252	2	45	*vly*, *sld*	Vagina
GV37	GCA_001953155.1	Complete	1.7467	41.8	1,359	2	45	*vly*	Blood culture
HMP9231	GCA_000213955.1	Complete	1.7265	41.2	1,354	2	45	*vly*	Endometrium
FDAARGOS_568	GCA_003812765.1	Complete	1.7166	41.3	1,368	2	45	*vly*, *sld*	NA[a]
ATCC 14019	GCA_000159155.2	Complete	1.6674	41.4	1,209	2	45	*vly*, *sld*	Vaginal
409-05	GCA_000025205.1	Complete	1.6176	42.0	1,237	2	45	*vly*	Vaginal
UGENT 06.41	GCA_003293675.1	Complete	1.5635	42.1	1,162	2	45	*vly*	Vagina

The *G. vaginalis* genome sequences considered in this study ranged in size from 1.47 Mb (UGent 09.48) to 1.77 Mb (FDAARGOS_296), with an average of 1,241 coding DNA sequences (CDS). Furthermore, these genomes had between 1 and 3 rRNA loci, and the number of tRNA genes ranged from 44 to 47. These results were

consistent with those of the genomes of nonbifidobacterial taxa of the *Bifidobacteriaceae*family, exhibiting averages of 1,502 CDS and 2.6 rRNA operons per genome and numbers of tRNA genes ranging from 45 to 48.

Specifically, a statistical comparison between *G. vaginalis* chromosomes and nonbifidobacterial genomes showed that, within this latter group, the average numbers of CDS, rRNA operons, and tRNA genes were increased by 17.41% ($P < 0.05$), 40.10% ($P < 0.05$), and 2.53% ($P < 0.05$), respectively. Moreover, the analogous comparison with members of the *Bifidobacterium* genus showed averages of 1,865 CDS and 3.2 rRNA loci and a tRNA content ranging from 40 to 79, revealing that within the bifidobacterial group, the averages of these numbers were increased by 33.45% ($P < 0.05$), 50.37% ($P < 0.05$), and 14.99% ($P < 0.05$), respectively (27). Based on the statistical comparisons of the number of CDS, rRNA, and tRNA, it seems at first that the *G. vaginalis* species has undergone a selective pressure similar to nonbifidobacterial members of the *Bifidobacteriaceae* family rather than members of the *Bifidobacterium* genus.

As mentioned above, *G. vaginalis* was correlated with BV incidence; nevertheless, it was also often found in healthy vaginal microbiota. It was supposed that certain lineages or species of *Gardnerella* are natural commensals and others can act as pathogens, triggering cases of symptomatic vaginal dysbiosis (29). To evaluate this hypothesis, we assessed the distribution of the two most studied and described genes that participate in the pathogenesis mechanism driven by *G. vaginalis*, i.e., those encoding the pore-forming CDC toxin vaginolysin (*vly*) and sialidase (*sld*), also known as neuraminidase (15, 17). Results showed that the genomes of 67 strains contained the *vly* gene, whereas more than half of the total number of *Gardnerella* chromosomes (40 genomes) were shown to encode a sialidase enzyme (Table 1). Notably, the sialidase enzymatic activity can reduce the protective

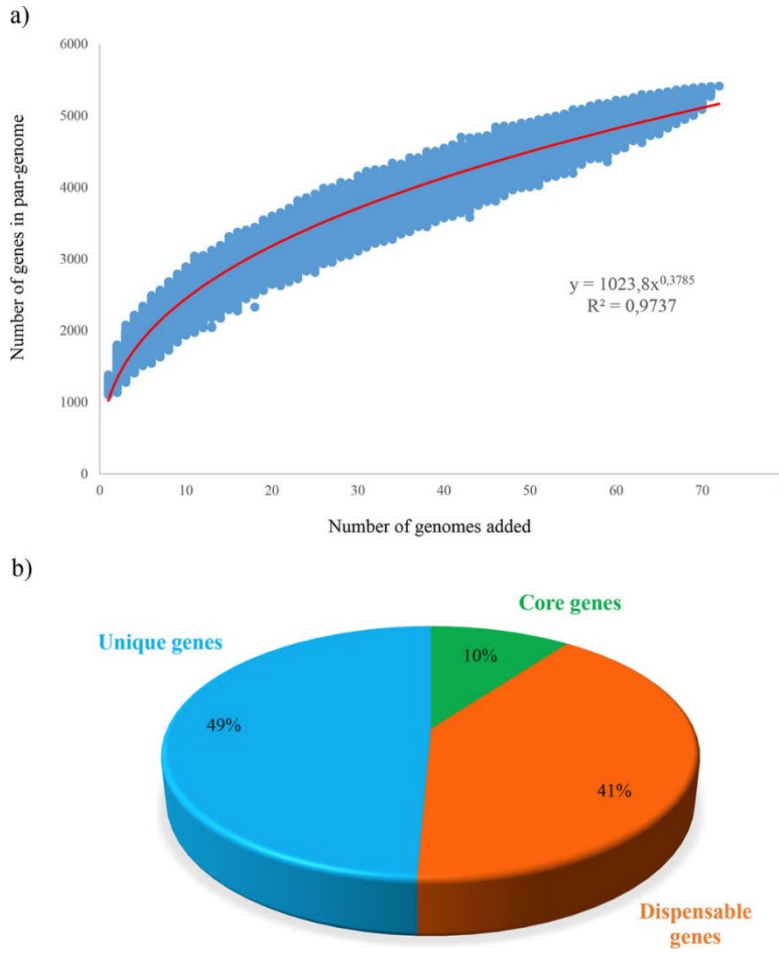
vaginal mucosal layer, facilitating bacterial adhesion to the vaginal epithelium and subsequent microbial biofilm development, thus increasing the infectious capabilities of *G. vaginalis* strains (15).

The evaluation of the possible presence of mobile elements within *G. vaginalis* chromosomes, followed by investigations of the genomic regions adjacent to both their ends, allowed us to assess the occurrence of eight putative virulence genes in eight *G. vaginalis* genomes. Each of these protein-encoding genes contained a domain resembling the coding region for a virulence-related protein belonging to a member of the *Streptococcus* genus and was found alongside a putative genomic prophage island (Fig. S1 in the supplemental material). Furthermore, 15 strains of *G. vaginalis* contained noteworthy genes placed tightly adjacent to transposases predicted to belong to members of the IS256 and IS3 families. These genes included a sequence encoding a RelE/RelB toxin-antitoxin system that is thought to exert toxic effects on both bacterial and eukaryotic cell types (30), as well as a ribosomal protection protein (TetM) conferring tetracycline resistance (31) and a collagen-binding protein (Fig. S1). These characteristics may reflect how prophage-like sequences and insertion sequences (IS) elements can be responsible for genomic duplications, deletions, and rearrangements, contributing to the genetic makeup and biodiversity of this bacterial taxon (32).

Pan-genome and core genome of the *G. vaginalis* species

Previous comparative genomic studies involving much smaller numbers of *G. vaginalis* genome sequences highlighted significant genomic differences between the chromosomes of this species (19, 24, 33). In this context, pan-genome reconstruction can contribute to deciphering the evolutionary dynamics, i.e., selection pressure of beneficial genes, as well as species- and genus-level differences

in overall gene content (34). In order to explore genetic differences, the genomes of 72 *G. vaginalis* strains were submitted to gene reannotation and subsequently analyzed from a pan-genome perspective, also unveiling their core genome and unique gene sequences. The pan-genome size of *G. vaginalis* has been shown to consist of 5,071 clusters of orthologous groups (COGs), and plotting it on a logarithmic scale as a function of the total amount of involved genomes revealed that the power trend line had not yet reached a plateau (Fig. 1). More precisely, adding a new *G. vaginalis* genome is predicted to add about 38 or 39 new genes to the *G. vaginalis* pan-genome.



***G. vaginalis* pan-genome. (a)** Pan-genome represented as a variation in size of the gene pool resulting from the sequential addition of the 72 *G. vaginalis* genomes. **(b)** Pie chart of the number of core genes (green), dispensable genes (orange), and unique genes (light blue) of *G. vaginalis*.

As previously mentioned, the pan-genome analysis allowed the evaluation of the core genome, defined as the set of gene families shared by all the organisms (34). In this comparison, a total repertoire of 514 COGs (10.1%) has been identified as a constituent of the core genome of *G. vaginalis*. Previous pan-genome analysis, including 60 *Bifidobacterium pseudolongum* genomes with an average genome size of 2.01 Mb, revealed a pan-genome consisting of 6,172 COGs corresponding to a core genome of 1,069 COGs (17.3%) (35). Likewise, a pan-genome curve based on

33 *Bifidobacterium longum* genomes with an average genome size of 2.35 Mb showed a pan-genome consisting of about 6,000 COGs and a core genome formed by 1,145 COGs (about 19%) (36). This evidence suggests that the *G. vaginalis* core genome could be considered smaller than that of other species belonging to the closely related *Bifidobacterium* genus.

Furthermore, through pan-genome analysis, we also identified the truly unique genes (TUGs) of *Gardnerella*, which ranged from 7 for the strain UMB0386 to 143 for KA00225. These findings showed that this species displays a modestly sized core genome corresponding to a relatively sizeable dispensable genome, i.e., the subset of genes shared by two or more strains (Fig. 1). Afterward, in silicoanalysis employing the eggNOG database allowed us to investigate the functional annotation of core genes. Excluding 14.9% that have no function, the large part of the encoded proteins belonging to the core proteome of *G. vaginalis* was related to essential cell maintenance, including translation (16.2%), carbohydrates, amino acids, and nucleotide metabolic processes (7.3%, 6.8%, and 6.6%, respectively) as well as inorganic ion transport (6.4%) (Fig. 1).

In addition, to get insights into specific genes supporting the adaptation of *G. vaginalis* to the vaginal environment, the genes belonging exclusively to the core genome of this species were further analyzed. A collection of 379 COGs, constituting the specific core genome of *G. vaginalis*, were obtained from the total amount of 514 COGs following the exclusion of COGs shared with other members of the *Bifidobacteriaceae* family (see Materials and Methods). This set of genes was evaluated from a functional annotation perspective. Such analysis revealed the ubiquitous presence of genes encoding C69-family dipeptidase, previously recognized as responsible for collagen molecule degradation (37), and a pullulanase, which seems to allow the efficient utilization of glycogen, i.e., the primary available

carbon source in the vaginal lumen (38). The mere presence of the latter genes within the *G. vaginalis* chromosomes cannot demonstrate that these genes are still under selective pressure. Thus, further investigations are requested to confirm their activity and functionality. However, their presence in the genomes of *G. vaginalis* may represent a clue to genetic adaptation to the vaginal environment of this species.

The screening of *G. vaginalis* genomes revealed the presence of several common features related to virulence, i.e., cytotoxicity/hemolysis mechanisms, biofilm production, iron uptake, adhesion to the epithelium, and antimicrobial resistance. Specifically, the ability of *G. vaginalis* to adhere to the vaginal wall is mediated by genes encoding type IV Flp pili (Table S3). At the same time, the subsequent biofilm development seems to be related to type I glycosyltransferase, also involving sortase enzyme activity, which was detected in each *G. vaginalis* genome analyzed as well (39). Furthermore, *G. vaginalis* genomes contained genes associated with toxicity, including a CDC toxin, vaginolysin, highly conserved among *G. vaginalis* strains (17), and a serralysin characterized in Serratia marcescens annotated as serralysin (40). Finally, within the core *G. vaginalis* genes, seven genes encoding putative drug resistance proteins were found, including two genes that are predicted to confer resistance to macrolide antibiotics, one major facilitator superfamily (MFS) transporter, as well as four unknown multidrug efflux systems.

Phylogenomic analysis of *G. vaginalis* taxon

As previously mentioned, the *Gardnerella* genus is currently considered to be composed of just one species, *G. vaginalis* (41). Over time, since its discovery, *G. vaginalis* was renamed repeatedly. This complicated taxonomic classification history provides an idea of the difficult challenge faced due to considerable diversity within this species. In recent years, phylogenetic analysis based on comparison of

chaperonin-60 (cpn60) sequences identified four subgroups within 112 *G. vaginalis* isolates (42). In contrast, analyses employing the bacterial 16S rRNA gene sequences did not give reliable support for a species-level resolution. To date, clear species identification events and the resultant presence of different species within the *Gardnerella* genus remain undiscovered. In this context, genomic comparisons represent a powerful in silico approach to highlight genomic differences between *G. vaginalis* strains, also contributing to its taxonomic classification. To infer the possible existence of phylogenomic-based clades within this species, the set of genes representing the core genome of *G. vaginalis* species was employed to perform a phylogenomic comparison. Specifically, we computed a phylogenetic tree based on the concatenation of 334 amino acid sequences (Fig. 2). Selected orthologous sequences were collected for previous genome comparison of all the chromosomes of 72 strains of *G. vaginalis*, together with the genome sequences of *Scardovia inopinata* JCM 12537 as a representative outgroup (Fig. 2). Remarkably, the resulting tree showed that most of the 72 *G. vaginalis* strains were grouped in two main clusters sharing the same phylogenetic branch. Moreover, within each cluster, it was possible to identify two additional groups, suggesting the existence of four putative different Gardnerella taxa (Fig. 2). Interestingly, *G. vaginalis* KA00225 and *G. vaginalis* CMW7778B were placed on separate branches.

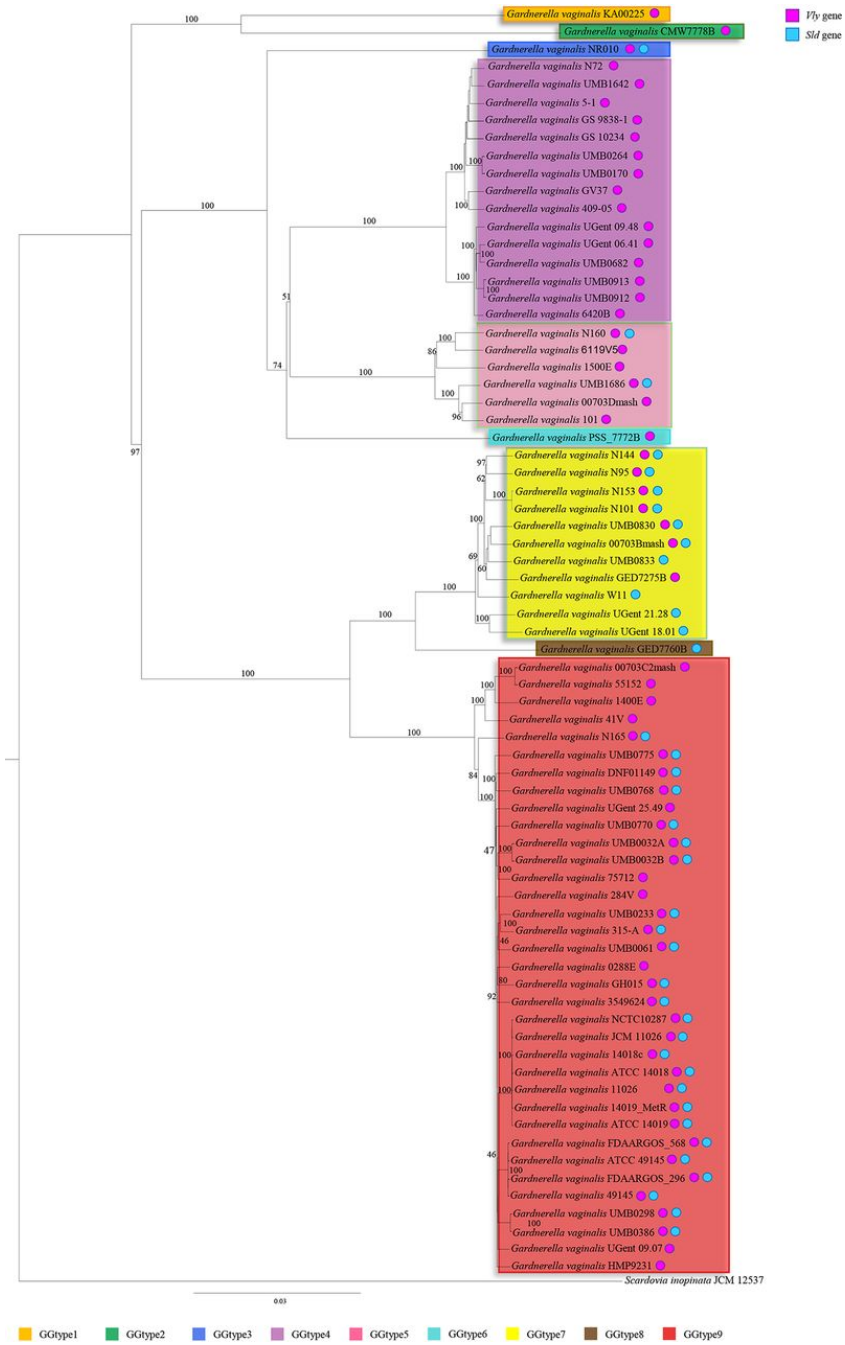


Figure 2. **Phylogenomic tree of *G. vaginalis*.** A proteomic tree was constructed based on the concatenation of 334 *G. vaginalis* core genes identified in the pan-genome analysis of the 72 *G. vaginalis* strains. The tree was built by the neighbor-joining method, and bootstrap percentages above 50 are shown at node points, based on 1,000 replicates. Phylogenetic clusters of different genotypes are highlighted in different colors. Colored circles represent the occurrence of the *vly* (dark pink) and *sld* (light blue) genes in the corresponding *G. vaginalis* genomes.

In order to further explore the genomic differences among members of the G. vaginalis taxon, the pairwise percent average nucleotide identity (ANI) was assessed, resulting in values ranging from 99.9% to 81.5% (Table S4). Notably, previous studies employing ANI analysis to taxonomically distinct species of the Bifidobacteriaceae family identified an ideal ANI threshold value of 94% (27, 43).

The analysis of ANI values among members of *G. vaginalis* revealed that the collected genome sequences fall into four main groups, within which ANI values were found above the species-level cutoff threshold of 94%. Conversely, genomes belonging to *G. vaginalis* strains GED7760B, PSS_7772B, CMW7778B, KA00225, and NR010 exhibited ANI values lower than 92% against each analyzed strain, highlighting another five putative different species of the *Gardnerella* genus (Table S4). These findings, together with data generated from the phylogenetic tree reconstruction, strongly support the existence of an extensive level of genomic variability between *G. vaginalis* strains and would cast doubt on the presence of a single species within this genus. Specifically, the calculation of ANI values allowed us to identify nine *Gardnerella* genotypes (GGtype1 to GGtype9), corresponding to putative different *Gardnerella* taxa (Fig. 2). Moreover, combining the genomic information related to the identified *G. vaginalis* virulence factors, we observed a heterogeneous distribution across the phylogenomic tree. In particular, the sialidase gene was detected almost exclusively in the genome sequences of *G. vaginalis* strains belonging to GGtype7 and GGtype9, together with the strains GED7760B and NR010, representative of GGtype8 and GGtype3, respectively (Fig. 2), suggesting these may be virulent genotypes. Conversely, the genomes of GGtype1, GGtype2, GGtype4, GGtype5, and GGtype6 appeared to lack the sialidase gene, revealing that these may be less virulent. Previous findings supported the existence of *G. vaginalis* strains that can provoke severe damage to the vaginal integrity through their ability to develop microbial biofilm and others that are linked with an asymptomatic medical

condition (19). Thus, our results highlighted how *G. vaginalis* strains encoding sialidase might be phylogenetically related, reinforcing the notion of a putative subdivision in potentially pathogenic and commensal strains (Fig. 2).

Assessing the prevalence and the abundance of *G. vaginalis* genotypes among the human population

Our genome-based analyses demonstrated that *G. vaginalis* species consists of separate subgroups. In light of the above findings, we assessed the composition of the vaginal microbiota of 175 women, aiming to investigate the prevalence and the distribution of the nine *G. vaginalis* genotypes among the human population. Specifically, a preliminary survey was performed to evaluate the overall vaginal microbiota composition of the collected 175 vaginal samples, displaying an abundance of *G. vaginalis* taxon above 5% in 20% of the samples (see Materials and Methods). Samples that did not reach such a threshold were discarded, resulting in a final collection of 34 metagenomic data sets, showing an abundance of *G. vaginalis* genomic reads ranging from 6.01% to 86.41%. Notably, six of the collected metagenomic data sets were obtained from vaginal samples of healthy pregnant women. In contrast, for the vast majority of the remaining 28 samples, it was not possible to get enough information regarding the health conditions of the subjects since the corresponding metadata was not available. Thereafter, these metagenomic data sets were assayed for the presence of the nine genotypes of *G. vaginalis* identified above, employing genome sequences belonging to strains KA00225, CMW7778B, NR010, UMB0264, 6119V5, PSS_7772B, 00703Bmash, GED7760B, and FDAARGOS_568 as representatives of each genotype. The minimum coverage of each gene was calculated based on the metagenomics reads with at least 99% full-length identity (see Materials and Methods).

As displayed in Fig. 3, considering the uneven distribution and abundance of the nine *G. vaginalis* genotypes, it was possible to delineate four groups overall within the collected vaginal samples. In particular, *G. vaginalis* communities with a predominance of a single genotype were identified within 16 metagenomic data sets. More specifically, the latter showed a predominance of GGtype4 in group C (n = 10) and GGtype3 in group B (n = 6), with an average percentage of metagenomic reads of 66.03% and 74.91%, respectively. Moreover, group A (n = 5) was mainly constituted by a combination of the latter two genotypes together with GGtype9 (Fig. 3). Interestingly, GGtype9 and GGtype3 were identified as putative virulence genotypes; thus, their presence in the vaginal microbiota could be linked with possible adverse health effects (Fig. 2). Conversely, GGtype4, which was found dominant in group C, seems to have less virulence potential since it does not contain the sialidase gene.

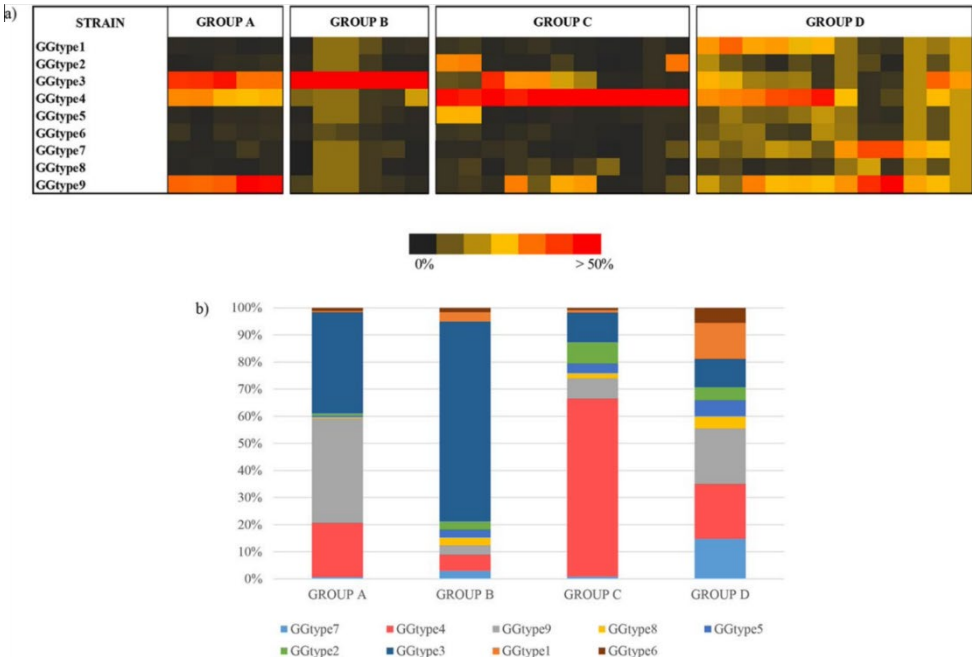


Figure 3 Metagenomic abundance of the different *G. vaginalis* genotypes. (a) Prevalence and distribution of the nine *G. vaginalis* genotypes observed in the 34 metagenome vaginal samples. **(b)** Average abundance of reads of each *G. vaginalis* genotype in the individuated four groups.

These findings allowed us to observe that different genotypes can be found within the female population, postulating their different impact on the vaginal environment. Moreover, these results may be consistent with the notion that *G. vaginalis* species can lead to a symptomatic unbalanced state of the vaginal microbiota in some instances and behave as natural commensal in others (18). Nevertheless, the involvement of specific genotypes in the development of significant clinical conditions should be investigated more in-depth in future metagenomic analyses that also include BV-positive vaginal microbiota samples.

Notably, all the metagenomic data sets from pregnant women fall in the same group (group D), characterized by the absence of a single predominant genotype. In fact, our results showed that the vaginal microbiota of pregnant women harbors a greater *G. vaginalis* biodiversity than that typical of nonpregnant women. It is known

that during the late gestational period, the microbiome undergoes significant strain-level variation, and the physiological state of pregnancy may have an impact also on the structure of *G. vaginalis* communities (44).

Phylogenomic evaluation of G. vaginalis within *Bifidobacteriaceae*

To date, *G. vaginalis* is considered a member of the *Bifidobacteriaceae* family since close relationships between this species and *Bifidobacterium* spp., based on 16S rRNA gene sequencing, were observed (45). Aiming to investigate the positioning of *G. vaginalis* within the *Bifidobacteriaceae* family, we performed a further phylogenetic analysis, including one representative strain for each genotype of *G. vaginalis* identified above by means of ANI values calculation, i.e., strains KA00225, CMW7778B, NR010, UMB0264, 6119V5, PSS_7772B, 00703Bmash, GED7760B, and FDAARGOS_568. These strains, together with the 96 type strains of the *Bifidobacteriaceae* family (bifidobacterial as well as nonbifidobacterial taxa) and *Cutibacterium acnes* KPA171202 as an outgroup, were employed to perform a comparative genomics analysis aimed to identify ubiquitously conserved protein sequences. The concatenation of 91 amino acid sequences shared between all considered genomes was used to construct the phylogenetic tree of the *Bifidobacteriaceae* family (Fig. 4). This analysis showed that most of the nodes were supported by 100% of the bootstrap values, validating the reliability of the phylogenetic tracing and robustness of the results. In accordance with a previous study, the obtained tree showed that *Bifidobacterium* spp. are separated from nonbifidobacterial taxa belonging to the genera *Scardovia*, *Parascardovia*, and *Alloscardovia* (27). Furthermore, these latter represent the deepest branches of the *Bifidobacteriaceae* family tree and, therefore, evidence of a very early separation in the evolution of this family. Focusing on the *G. vaginalis* genotypes, this

phylogenetic investigation highlighted its evolutionary positioning within the *Bifidobacterium* genus.

More specifically, *Bifidobacterium tsurumiense* was identified as the phylogenetically closest-related taxon to *G. vaginalis*. Furthermore, the type strain *G. vaginalis* ATCC 14019 exhibits tight phylogenetic grouping with strains belonging to GGtype9, which is consistent with our ANI-based findings (see above). Interestingly, the nine *G. vaginalis* strains representative of the many related putative species are grouped, giving rise to a new cluster located alongside the previously described *Bifidobacterium boum* group (46). Overall, these findings clearly showed that by employing a robust phylogenomic-based approach, the *G. vaginalis* species resulted in being identified along with some currently classified *Bifidobacterium* species, thus suggesting the need for a reevaluation of the currently known taxonomy of the *Bifidobacterium* genus.

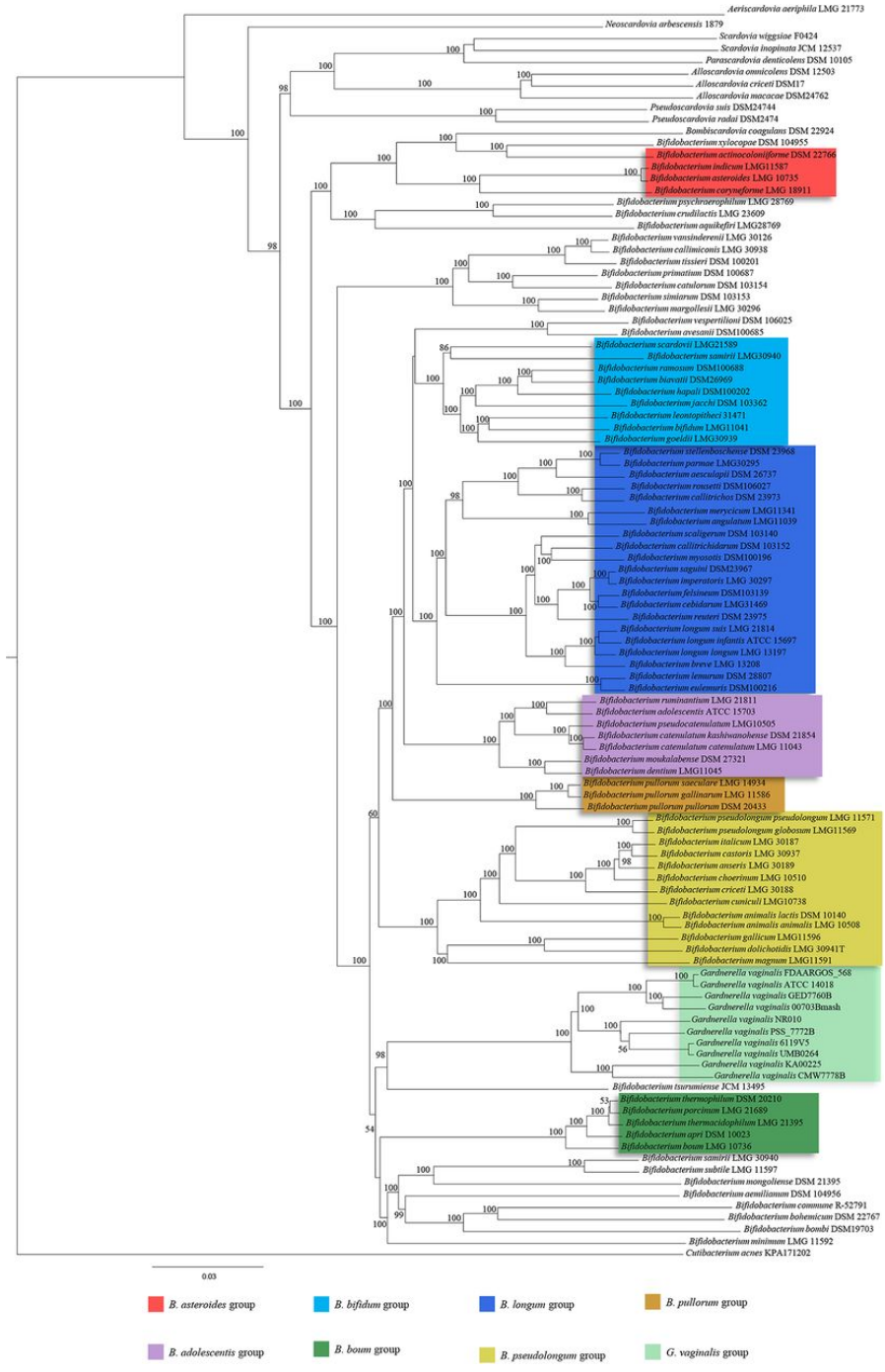


Figure 4. Phylogenomic tree of the *Bifidobacteriaceae* family. The proteomic tree is based on the concatenation of 91 core genes shared by members of the *Bifidobacteriaceae* family. The tree was constructed by the neighbor-joining method, and bootstrap percentages above 50 are shown at node points, based on

1,000 replicates. Phylogenetic groups are highlighted in different colors. The *G. vaginalis* cluster is highlighted in light green.

In conclusion, the high degree of genetic heterogeneity observed among members of *Gardnerella vaginalis* has been investigated, suggesting inaccuracy in the current taxonomic classification that consists of a single species within the *Gardnerella* genus.

In this study, through an exhaustive phylogenomic and comparative genomic analysis employing 72 publicly available *G. vaginalis* genome sequences, we identified nine different Gardnerella genotypes (GGtype1 to GGtype9). Notably, within the Bifidobacteriaceae family, *G. vaginalis* is phylogenetically located alongside the *Bifidobacterium boum* group (46), casting doubt on its current taxonomic classification due to the relatedness with other bifidobacterial species. Furthermore, the characterization of the pan-genome of *G. vaginalis* allowed us to obtain insights into the adaptation mechanisms to the vaginal environment.

Our data showed that genes encoding collagen and glycogen utilization functions were ubiquitous genetic elements, while virulence-associated ones exhibited an uneven distribution among genotypes. Notably, among *G. vaginalis* genes encoding virulence factor, sialidase is especially noteworthy due to its involvement in the degradation of the vaginal mucosal layer as well as microbial biofilm formation (15). The latter gene was identified between members of four genotypes, i.e., GGtype3, GGtype7, GGtype8, as well as GGtype9, allowing to discern those genotypes with the highest putative virulence capability and potentially linked with major adverse health outcomes. Interestingly, the microbial profiling of the vaginal microbiota of 34 women allowed us to identify GGtype3 and GGtype9 as sialidase positive, as well as GGtype4, which conversely lacks the sialidase gene, as the most abundant

genotypes among the human population. These findings are in line with previous studies since both pathogenic and commensal *G. vaginalis* strains have been previously described (11).

MATERIALS AND METHODS

Gardnerella vaginalis and *Bifidobacteriaceae* genome sequences

Genome sequences of *G. vaginalis* strains were retrieved from the National Center for Biotechnology Information (NCBI) public database, resulting in 107 available genomes. Moreover, incomplete genomes (genome size less than 1.4 Mb) as well as genome sequences that exhibited low sequencing quality (genome coverage lower than 30× or containing unspecified nucleotide bases in conformity to IUPAC nomenclature), were discarded. Furthermore, a comparison of the *G. vaginalis* genome sequences was performed to evaluate the average nucleotide identity (ANI) values for each genome with respect to the genome of *G. vaginalis* ATCC 14018, which is the type strain of this species. Based on this analysis, there were inconsistencies in the predicted taxonomy of two strains belonging to the *Lactobacillus* genus, i.e., *G. vaginalis* UMB0388 and *G. vaginalis* MGYG-HGUT-00021. Finally, collected high-quality genome sequences of 72 *G. vaginalis* (Table 1) were compared to each other. Additional genomic and phylogenomic analyses were performed employing 96 type strains of the *Bifidobacteriaceae* family retrieved from the NCBI database, including 84 bifidobacterial genome sequences and 12 nonbifidobacterial genome sequences (27, 46) (Table S1 in the supplemental material).

Genome annotation

In order to obtain comparable quality standards for the analyzed genomes, the 72 *G. vaginalis* genome sequences retrieved from the NCBI database were submitted to

annotation employing the MEGAnnotator pipeline (47). Protein-encoding open reading frames (ORFs) were predicted using Prodigal (48). tRNA genes were detected using tRNAscan-SE v1.4 (49), while rRNA genes were identified using RNAmmer v1.2 (50). Outcomes of the gene-finder program were combined with data from RAPSearch2 analysis (Reduced Alphabet based Protein similarity Search) (51) of a nonredundant protein database provided by the NCBI and hidden Markov model profile (HMM) search (http://hmmer.org/) in the manually curated Pfam-A protein family database (52). Results were examined by Artemis (53), which was used for validating predicted genes and, where required, for genome manual editing consisting of removal or addition of coding regions as well as a redefinition of gene starts.

Virulence gene identification

In order to perform a screening among genomes of the 72 *G. vaginalis* strains, amino acid sequences of nonredundant WP accessions, i.e., a CDC vaginolysin and 26 exo-alpha-sialidases, were retrieved from the Identical Protein Groups (IPG) resource of the NCBI database (https://www.ncbi.nlm.nih.gov/ipg/). Then, putative CDC vaginolysin and sialidase genes were identified through BLASTP analysis (E value cutoff of $1E^{-5}$) (54). A subsequent manual inspection of the resulting aligned proteins based on their amino acid sequence identity (greater than 42%), combined with the alignment length (more than 500 amino acids), allowed us to discard false positives from the prediction (Table 1). In addition, careful scrutiny of *G. vaginalis* core genes and the genomic regions adjacent to both ends of the identified mobile elements allowed us to discover further virulence traits. Such outcomes were subsequentially validated through the cross-examination of the virulence factor database (VFDB) (55).

Prophages and IS element identification

The 72 *G. vaginalis* genomes were screened for prophage-associated genes using a custom database employing BLASTP analysis (54) (E value cutoff of $1E^{-5}$). The custom database was assembled through previously bifidoprophage-validated sequences retrieved from 60 bifidoprophages previously described (56). Then, a manual examination of the DNA region surrounding a putative prophage-encoding gene was performed, allowing the identification of complete prophage-like sequences (Table S2). Moreover, the same *G. vaginalis* genomes were also screened for the presence of IS elements (57) through the IS Finder online tool (https://isfinder.biotoul.fr/) (Table S2).

***G. vaginalis* pan-genome analysis**

A pan-genome calculation employing 72 genomes of *G. vaginalis* was performed using the PGAP (pan-genome analysis pipeline) (58). Predicted ORFs were organized into functional clusters employing the GF (gene family) method, which consists of a similarity search between each protein pair through BLAST analysis (cutoff E value of 1×10^{-10} and 50% identity over at least 80% of both protein sequences). Following this, a clustering in protein families of orthologous genes was performed using MCL (graph theory-based Markov clustering algorithm) (59). A pan-genome profile was built using an optimized algorithm integrated into PGAP software, based on a presence/absence matrix that included all protein families of orthologous genes identified in the analyzed genomes. Subsequently, the unique protein families for each of 72 *G. vaginalis* genomes were identified. Protein families shared between all genomes allowed us to build the core genome of the *G. vaginalis* species, defined by selecting the families that contained at least one protein member for each genome. A different pan- and core- genome analysis was performed on the 96 *Bifidobacteriaceae* type strains as described above, including *G.*

vaginalis ATCC 14018, identifying 135 COGs belonging to the core genome of this family. Afterward, in order to obtain the core genes of *G. vaginalis* that were not shared with other members of the *Bifidobacteriaceae*family, the 135 gene sequences attributed to *G. vaginalis* ATCC 14018 were used to remove the corresponding COGs from the core genome of *G. vaginalis* species.

Phylogenomic comparison between *G. vaginalis* strains and their positioning within the Bifidobacteriaceae family

In order to assess genome differences between *G. vaginalis* strains, a phylogenetic comparison involving the 72 genome sequences retrieved from NCBI was performed. For this purpose, the concatenated core genome sequences were aligned using MAFFT (60), and the resulting phylogenetic tree was constructed using the neighbor-joining method in Clustal W v2.1 (61). A visual core genome tree was developed using FigTree software (http://tree.bio.ed.ac.uk/software/figtree/). A value for the average nucleotide identity (ANI) was calculated for each genome pair using the fastANI software (62).

A further phylogenomic analysis, aiming to evaluate the phylogenetic position of *G. vaginalis* within the family *Bifidobacteriaceae*, was executed on 72 *G. vaginalis* genome sequences together with 96 *Bifidobacteriaceae* type strains as described above.

Whole-genome sequencing data collection and analysis

The publicly available vaginal metagenomic data sets were retrieved from NCBI (BioProject accession no. PRJEB24147, PRJNA352475, PRJNA361427, PRJNA576566, and PRJNA379120). Specifically, we selected Illumina whole-genome shotgun (WGS) sequencing data concerning vaginal samples from midvagina and cervix swabs of fertile pregnant, as well as nonpregnant, women. The

resulting 175 vaginal metagenomic data sets were analyzed through a shallow shotgun metagenomics approach (63), allowing us to achieve high taxonomic resolution at the species level. In order to reconstruct the microbiota composition of vaginal samples, the fastq files of the paired-end reads were used as input for the genome assemblies through the METAnnotatorX pipeline (64). The SPAdes software was used for *de novo* assembly of each genome sequence (65). To assess the distribution of the different *G. vaginalis* genotypes among the human population, the samples showing a relative abundance of this species below 5% were discarded. In fact, it was observed that below this threshold level, the number of *G. vaginalis* reads within samples was not enough to ensure a reasonable mapping accuracy of genotypes (see below). Afterward, the genome sequences belonging to the nine strains representative of as many *G. vaginalis* genotypes identified in this study were aligned with WGS reads. Metagenomics data sets were filtered by use of the fastq-mcf script (https://expressionanalysis.github.io/ea-utils/) (minimum mean quality score, 20; window size, 5 bp; quality threshold, 25; and minimum length, 80 bp) to obtain high-quality reads. Collected reads were aligned against the human genome using the Burrows-Wheeler Aligner program (66) (BWA-MEM algorithm with trigger reseeding, 1.5; minimum seed length, 19; matching score, 1; mismatch penalty, 4; gap open penalty, 6; and gap extension penalty, 1) and processed with the SAMtools software package (67), aiming to remove human reads. The final mapping against the genome sequences of the *G. vaginalis* genotypes was performed using Bowtie 2 (68) through multiple-hit mapping and "very-sensitive" policy. The mapping was performed using a minimum score threshold function (–score-min C,−13,0) to limit reads of arbitrary length to two mismatches and retain those matches with at least 99% full-length identity. HTSeq software (69) (running in union mode) was employed to calculate read counts corresponding to the *G. vaginalis* genotypes.

ACKNOWLEDGMENTS

We thank GenProbio Srl for the financial support of the Laboratory of Probiogenomics.

Part of this research was conducted using the High Performance Computing (HPC) facility of the University of Parma.

References

1. Bradford LL, Ravel J. 2017. The vaginal mycobiome: a contemporary perspective on fungi in women's health and diseases. Virulence 8:342–351. doi: 10.1080/21505594.2016.1237332.

2. Taha TE, Hoover DR, Dallabetta GA, Kumwenda NI, Mtimavalye LA, Yang LP, Liomba GN, Broadhead RL, Chiphangwi JD, Miotti PG. 1998. Bacterial vaginosis and disturbances of vaginal flora: association with increased acquisition of HIV. AIDS 12:1699–1706. doi: 10.1097/00002030-199813000-00019.

3. Sobel JD. 1999. Is there a protective role for vaginal flora? Curr Infect Dis Rep 1:379–383. doi: 10.1007/s11908-999-0045-z.

4. Wiesenfeld HC, Hillier SL, Krohn MA, Landers DV, Sweet RL. 2003. Bacterial vaginosis is a strong predictor of Neisseria gonorrhoeae and Chlamydia trachomatis infection. Clin Infect Dis 36:663–668. doi: 10.1086/367658.

5. Skarin A, Sylwan J. 1986. Vaginal lactobacilli inhibiting growth of Gardnerella vaginalis, Mobiluncus and other bacterial species cultured from vaginal content of women with bacterial vaginosis. Acta Pathol Microbiol Immunol Scand B 94:399–403. doi: 10.1111/j.1699-0463.1986.tb03074.x.

6. Nam H, Whang K, Lee Y. 2007. Analysis of vaginal lactic acid producing bacteria in healthy women. J Microbiol 45:515–520.

7. Oh HY, Kim BS, Seo SS, Kong JS, Lee JK, Park SY, Hong KM, Kim HK, Kim MK. 2015. The association of uterine cervical microbiota with an increased risk for cervical intraepithelial neoplasia in Korea. Clin Microbiol Infect 21:674.e1–674.e9. doi: 10.1016/j.cmi.2015.02.026.

8. Lee JE, Lee S, Lee H, Song YM, Lee K, Han MJ, Sung J, Ko G. 2013. Association of the vaginal microbiota with human papillomavirus infection in a Korean twin cohort. PLoS One 8:e63514. doi: 10.1371/journal.pone.0063514.

9. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, Karlebach S, Gorle R, Russell J, Tacket CO, Brotman RM, Davis CC, Ault K, Peralta L, Forney LJ. 2011. Vaginal

microbiome of reproductive-age women. Proc Natl Acad Sci U S A 108(Suppl 1):4680–4687. doi: 10.1073/pnas.1002611107.

10. Gardner HL, Dukes CD. 1955. Haemophilus vaginalis vaginitis: a newly defined specific infection previously classified “nonspecific” vaginitis. Am J Obstet Gynecol 69:962–976. doi: 10.1016/0002-9378(55)90095-8.

11. Aroutcheva AA, Simoes JA, Behbakht K, Faro S. 2001. Gardnerella vaginalis isolated from patients with bacterial vaginosis and from patients with healthy vaginal ecosystems. Clin Infect Dis 33:1022–1027. doi: 10.1086/323030.

12. Ravel J, Brotman RM, Gajer P, Ma B, Nandy M, Fadrosh DW, Sakamoto J, Koenig SS, Fu L, Zhou X, Hickey RJ, Schwebke JR, Forney LJ. 2013. Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis. Microbiome 1:29. doi: 10.1186/2049-2618-1-29.

13. Hardy L, Jespers V, Dahchour N, Mwambarangwe L, Musengamana V, Vaneechoutte M, Crucitti T. 2015. Unravelling the bacterial vaginosis-associated biofilm: a multiplex Gardnerella vaginalis and Atopobium vaginae fluorescence in situ hybridization assay using peptide nucleic acid probes. PLoS One 10:e0136658. doi: 10.1371/journal.pone.0136658.

14. Tsang A, Bradbury JM. 1981. Separation and properties of prestalk and prespore cells of Dictyostelium discoideum. Exp Cell Res 132:433–441. doi: 10.1016/0014-4827(81)90118-X.

15. Hardy L, Jespers V, Van den Bulck M, Buyze J, Mwambarangwe L, Musengamana V, Vaneechoutte M, Crucitti T. 2017. The presence of the putative Gardnerella vaginalis sialidase A gene in vaginal specimens is associated with bacterial vaginosis biofilm. PLoS One 12:e0172522. doi: 10.1371/journal.pone.0172522.

16. Soong G, Muir A, Gomez MI, Waks J, Reddy B, Planet P, Singh PK, Kaneko Y, Kanetko Y, Wolfgang MC, Hsiao Y-S, Tong L, Prince A. 2006. Bacterial neuraminidase facilitates mucosal infection by participating in biofilm production. J Clin Invest 116:2297–2305. doi: 10.1172/JCI27920.

17. Gelber SE, Aguilar JL, Lewis KL, Ratner AJ. 2008. Functional and phylogenetic characterization of vaginolysin, the human-specific cytolysin from Gardnerella vaginalis. J Bacteriol 190:3896–3903. doi: 10.1128/JB.01965-07.

18. Janulaitiene M, Paliulyte V, Grinceviciene S, Zakareviciene J, Vladisauskiene A, Marcinkute A, Pleckaityte M. 2017. Prevalence and distribution of Gardnerella vaginalis subgroups in women with and without bacterial vaginosis. BMC Infect Dis 17:394. doi: 10.1186/s12879-017-2501-y.

19. Cornejo OE, Hickey RJ, Suzuki H, Forney LJ. 2018. Focusing the diversity of Gardnerella vaginalis through the lens of ecotypes. Evol Appl 11:312–324. doi: 10.1111/eva.12555.

20. Harwich MD Jr, Alves JM, Buck GA, Strauss JF III, Patterson JL, Oki AT, Girerd PH, Jefferson KK. 2010. Drawing the line between commensal and pathogenic Gardnerella vaginalis through genome analysis and virulence studies. BMC Genomics 11:375. doi: 10.1186/1471-2164-11-375.

21. Zinnemann K, Turner GC. 1963. The taxonomic position of “Haemophilus vaginalis” [Corynebacterium vaginale]. J Pathol 85:213–219. doi: 10.1002/path.1700850120. [CrossRef] [Google Scholar]

22. Pickett J. 1980. Transfer of Haemophilus vaginalis Gardner and Dukes to a new genus, Gardnerella: G. vaginalis (Gardner and Dukes) comb. nov. Int J Syst Evol Microbiol 30. doi: 10.1099/00207713-30-1-170. [CrossRef] [Google Scholar]

23. Piot P, van Dyck E, Goodfellow M, Falkow S. 1980. A taxonomic study of Gardnerella vaginalis (Haemophilus vaginalis) Gardner and Dukes 1955. J Gen Microbiol 119:373–396. doi: 10.1099/00221287-119-2-373.

24. Ahmed A, Earl J, Retchless A, Hillier SL, Rabe LK, Cherpes TL, Powell E, Janto B, Eutsey R, Hiller NL, Boissy R, Dahlgren ME, Hall BG, Costerton JW, Post JC, Hu FZ, Ehrlich GD. 2012. Comparative genomic analyses of 17 clinical isolates of Gardnerella vaginalis provide evidence of multiple genetically isolated clades consistent with subspeciation into genovars. J Bacteriol 194:3922–3937. doi: 10.1128/JB.00056-12.

25. Pleckaityte M, Janulaitiene M, Lasickiene R, Zvirbliene A. 2012. Genetic and biochemical diversity of Gardnerella vaginalis strains isolated from women with bacterial vaginosis. FEMS Immunol Med Microbiol 65:69–77. doi: 10.1111/j.1574-695X.2012.00940.x.

26. Vaneechoutte M, Guschin A, Van Simaey L, Gansemans Y, Van Nieuwerburgh F, Cools P. 2019. Emended description of Gardnerella vaginalis and description of Gardnerella leopoldii sp. nov., Gardnerella piotii sp. nov. and Gardnerella swidsinskii sp. nov., with delineation of 13 genomic species within the genus Gardnerella. Int J Syst Evol Microbiol 69:679–687. doi: 10.1099/ijsem.0.003200.

27. Lugli GA, Milani C, Turroni F, Duranti S, Mancabelli L, Mangifesta M, Ferrario C, Modesto M, Mattarelli P, Jiří K, van Sinderen D, Ventura M. 2017. Comparative genomic and phylogenomic analyses of the Bifidobacteriaceae family. BMC Genomics 18:568. doi: 10.1186/s12864-017-3955-4.

28. Dutta C, Paul S. 2012. Microbial lifestyle and genome signatures. Curr Genomics 13:153–162. doi: 10.2174/138920212799860698.

29. Schellenberg JJ, Patterson MH, Hill JE. 2017. Gardnerella vaginalis diversity and ecology in relation to vaginal symptoms. Res Microbiol 168:837–844. doi: 10.1016/j.resmic.2017.02.011.

30. Yamamoto TA, Gerdes K, Tunnacliffe A. 2002. Bacterial toxin RelE induces apoptosis in human cells. FEBS Lett 519:191–194. doi: 10.1016/S0014-5793(02)02764-3.

31. Grossman TH. 2016. Tetracycline antibiotics and resistance. Cold Spring Harb Perspect Med 6:a025387. doi: 10.1101/cshperspect.a025387.

32. Touchon M, Rocha EP. 2007. Causes of insertion sequences abundance in prokaryotic genomes. Mol Biol Evol 24:969–981. doi: 10.1093/molbev/msm014.

33. Santiago GL, Deschaght P, El Aila N, Kiama TN, Verstraelen H, Jefferson KK, Temmerman M, Vaneechoutte M. 2011. Gardnerella vaginalis comprises three distinct genotypes of which only two produce sialidase. Am J Obstet Gynecol 204:450 e1-7. doi: 10.1016/j.ajog.2010.12.061.

34. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-genome. Curr Opin Genet Dev 15:589–594. doi: 10.1016/j.gde.2005.09.006.

35. Lugli GA, Duranti S, Albert K, Mancabelli L, Napoli S, Viappiani A, Anzalone R, Longhi G, Milani C, Turroni F, Alessandri G, Sela DA, van Sinderen D, Ventura M. 2019. Unveiling genomic diversity among members of the species Bifidobacterium pseudolongum, a widely distributed gut commensal of the animal kingdom. Appl Environ Microbiol 85:e03065-18. doi: 10.1128/AEM.03065-18.

36. O'Callaghan A, Bottacini F, O'Connell Motherway M, van Sinderen D. 2015. Pangenome analysis of Bifidobacterium longum and site-directed mutagenesis through by-pass of restriction-modification systems. BMC Genomics 16:832. doi: 10.1186/s12864-015-1968-4.

37. Sakamoto T, Otokawa T, Kono R, Shigeri Y, Watanabe K. 2013. A C69-family cysteine dipeptidase from Lactobacillus farciminis JCM1097 possesses strong Gly-Pro hydrolytic activity. J Biochem 154:419–427. doi: 10.1093/jb/mvt069.

38. van der Veer C, Hertzberger RY, Bruisten SM, Tytgat HLP, Swanenburg J, de Kat Angelino-Bart A, Schuren F, Molenaar D, Reid G, de Vries H, Kort R. 2019. Comparative genomics of human Lactobacillus crispatus isolates reveals genes for glycosylation and glycogen degradation: implications for in vivo dominance of the vaginal microbiota. Microbiome 7:49. doi: 10.1186/s40168-019-0667-9.

39. Yeoman CJ, Yildirim S, Thomas SM, Durkin AS, Torralba M, Sutton G, Buhay CJ, Ding Y, Dugan-Rocha SP, Muzny DM, Qin X, Gibbs RA, Leigh SR, Stumpf R, White BA, Highlander SK, Nelson KE, Wilson BA. 2010. Comparative genomics of Gardnerella

vaginalis strains reveals substantial differences in metabolic and virulence potential. PLoS One 5:e12411. doi: 10.1371/journal.pone.0012411.

40. Zhang L, Morrison AJ, Thibodeau PH. 2015. Interdomain contacts and the stability of serralysin protease from Serratia marcescens. PLoS One 10:e0138419. doi: 10.1371/journal.pone.0138419.

41. Castro J, Jefferson KK, Cerca N. 2020. Genetic heterogeneity and taxonomic diversity among Gardnerella species. Trends Microbiol 28:202–211. doi: 10.1016/j.tim.2019.10.002.

42. Schellenberg JJ, Paramel Jayaprakash T, Withana Gamage N, Patterson MH, Vaneechoutte M, Hill JE. 2016. Gardnerella vaginalis subgroups defined by cpn60 sequencing and sialidase activity in isolates from Canada, Belgium and Kenya. PLoS One 11:e0146510. doi: 10.1371/journal.pone.0146510.

43. Lugli GA, Milani C, Turroni F, Duranti S, Ferrario C, Viappiani A, Mancabelli L, Mangifesta M, Taminiau B, Delcenserie V, van Sinderen D, Ventura M. 2014. Investigation of the evolutionary development of the genus Bifidobacterium by comparative genomics. Appl Environ Microbiol 80:6383–6394. doi: 10.1128/AEM.02004-14.

44. Goltsman DSA, Sun CL, Proctor DM, DiGiulio DB, Robaczewska A, Thomas BC, Shaw GM, Stevenson DK, Holmes SP, Banfield JF, Relman DA. 2018. Metagenomic analysis with strain-level resolution reveals fine-scale variation in the human pregnancy microbiome. Genome Res 28:1467–1480. doi: 10.1101/gr.236000.118.

45. Leblond-Bourget N, Philippe H, Mangin I, Decaris B. 1996. 16S rRNA and 16S to 23S internal transcribed spacer sequence analyses reveal inter- and intraspecific Bifidobacterium phylogeny. Int J Syst Bacteriol 46:102–111. doi: 10.1099/00207713-46-1-102.

46. Lugli GA, Milani C, Duranti S, Mancabelli L, Mangifesta M, Turroni F, Viappiani A, van Sinderen D, Ventura M. 2018. Tracking the taxonomy of the genus Bifidobacterium based on a phylogenomic approach. Appl Environ Microbiol 84:e02249-17. doi: 10.1128/aem.02249-17.

47. Lugli GA, Milani C, Mancabelli L, van Sinderen D, Ventura M. 2016. MEGAnnotator: a user-friendly pipeline for microbial genomes assembly and annotation. FEMS Microbiol Lett 363:fnw049. doi: 10.1093/femsle/fnw049.

48. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. doi: 10.1186/1471-2105-11-119.

49. Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964. doi: 10.1093/nar/25.5.955.

50. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35:3100–3108. doi: 10.1093/nar/gkm160.

51. Zhao Y, Tang H, Ye Y. 2012. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics 28:125–126. doi: 10.1093/bioinformatics/btr595.

52. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. 2014. Pfam: the protein families database. Nucleic Acids Res 42:D222–D230. doi: 10.1093/nar/gkt1223.

53. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. 2000. Artemis: sequence visualization and annotation. Bioinformatics 16:944–945. doi: 10.1093/bioinformatics/16.10.944.

54. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. doi: 10.1016/S0022-2836(05)80360-2.

55. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. 2005. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res 33:D325–D328. doi: 10.1093/nar/gki008.

56. Lugli GA, Milani C, Turroni F, Tremblay D, Ferrario C, Mancabelli L, Duranti S, Ward DV, Ossiprandi MC, Moineau S, van Sinderen D, Ventura M. 2016. Prophages of the genus Bifidobacterium as modulating agents of the infant gut microbiota. Environ Microbiol 18:2196–2213. doi: 10.1111/1462-2920.13154.

57. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res 34:D32–6. doi: 10.1093/nar/gkj014.

58. Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J. 2012. PGAP: pan-genomes analysis pipeline. Bioinformatics 28:416–418. doi: 10.1093/bioinformatics/btr655.

59. Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–1584. doi: 10.1093/nar/30.7.1575.

60. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066. doi: 10.1093/nar/gkf436.

61. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948. doi: 10.1093/bioinformatics/btm404.

62. Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. doi: 10.1038/s41467-018-07641-9.

63. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Gohl DM, Beckman KB, Knight R, Knights D. 2018. Evaluating the information content of shallow shotgun metagenomics. mSystems 3:e00069-18. doi: 10.1128/mSystems.00069-18.

64. Milani C, Casey E, Lugli GA, Moore R, Kaczorowska J, Feehily C, Mangifesta M, Mancabelli L, Duranti S, Turroni F, Bottacini F, Mahony J, Cotter PD, McAuliffe FM, van Sinderen D, Ventura M. 2018. Tracing mother-infant transmission of bacteriophages by means of a novel analytical tool for shotgun metagenomic datasets: METAnnotatorX. Microbiome 6:145. doi: 10.1186/s40168-018-0527-z.

65. Nurk S, Bankevich A, Antipov D, Gurevich AA, Korobeynikov A, Lapidus A, Prjibelski AD, Pyshkin A, Sirotkin A, Sirotkin Y, Stepanauskas R, Clingenpeel SR, Woyke T, McLean JS, Lasken R, Tesler G, Alekseyev MA, Pevzner PA. 2013. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. J Comput Biol 20:714–737. doi: 10.1089/cmb.2013.0084.

66. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. doi: 10.1093/bioinformatics/btp324.

67. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. doi: 10.1093/bioinformatics/btp352.

68. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. doi: 10.1038/nmeth.1923.

69. Anders S, Pyl PT, Huber W. 2015. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics 31:166–169. doi: 10.1093/bioinformatics/btu638

Chapter 7

Unraveling the Microbiome of Necrotizing Enterocolitis: Insights in Novel Microbial and Metabolomic Biomarkers

Chiara Tarracchini, Christian Milani, Giulia Longhi, Federico Fontana, Leonardo Mancabelli, Roberta Pintus, Gabriele Andrea Lugli, Giulia Alessandri, Rosaria Anzalone, Alice Viappiani, Francesca Turroni, Michele Mussap, Angelica Dessì, Flaminia Cesare Marincola, Antonio Noto, Anna De Magistris, Marine Vincent, Sergio Bernasconi, Jean-Charles Picaud, Vassilios Fanos #, Marco Ventura #

The results of this chapter were published in Microbiology Spectrum, 2021 Oct; doi: 10.1128/Spectrum.01176-21.

Abstract

Necrotizing enterocolitis (NEC) is among the most relevant gastrointestinal diseases affecting mostly prematurely born infants with low birth weight. While intestinal dysbiosis has been proposed as one of the possible factors involved in NEC pathogenesis, the role of the gut microbiota remains poorly understood. In this study, the gut microbiota of preterm infants was explored to highlight differences in the composition between infants affected by NEC and prior to NEC development. A large-scale gut microbiome analysis was performed, including 47 shotgun sequencing data generated in the framework of this study, along with 124 retrieved from publicly available repositories. Meta-analysis led to the identification of Preterm Community State Types (PT-CSTs), which recur in healthy controls and NEC infants. Such analyses revealed an overgrowth of a range of opportunistic microbial species accompanying the loss of gut microbial biodiversity in NEC subjects. Moreover, longitudinal insights into preterm infants prior to NEC development indicated *Clostridium neonatale* and *Clostridium perfringens* species as potential biomarkers for predictive early diagnosis of this disease. Furthermore, functional investigation of the enzymatic reaction profiles associated with pre-NEC condition suggested DL-lactate as a putative metabolic biomarker for early detection of NEC onset.

IMPORTANCE. Necrotizing enterocolitis (NEC) is a severe gastrointestinal disease predominantly occurring in premature infants whose etiology is still not fully understood. In this study, the analysis of infant fecal samples through shotgun metagenomics approaches revealed a marked reduction of the intestinal (bio)diversity and an overgrowth of (opportunistic) pathogens associated with the NEC development. In particular, dissection of the infant's gut microbiome before

NEC diagnosis highlighted the potential involvement of *Clostridium* genus members in the progression of NEC. Remarkably, our analyses highlighted a gastrointestinal DL-lactate accumulation among NEC patients that might represent a novel potential functional biomarker for the early diagnosis of NEC.

For Supplementary Materials see the article published in Microbiology Spectrum

INTRODUCTION

Necrotizing enterocolitis (NEC) is a harmful gastrointestinal disease commonly encountered in neonatal intensive care units (NICU) worldwide. Upon NEC occurrence, segments of the infant's gastrointestinal tract undergo ischemia and subsequently necrosis, thus representing a gastrointestinal emergency in neonatal age, occurring in about 8 % of premature infants with a reported mortality rate up to 25 % (1). NEC is believed to be a disease with a multifactor etiology whose precise cause has not been fully understood. However, several risk factors have been identified. In particular, premature birth (less than 32 weeks of gestation) and very low birth weight (< 1500 g) have been reported as amongst the main factors of increased risk of sepsis and NEC (1, 2). Besides, NEC has long been linked to microbial dysbiosis of the infant gut (3, 4). Indeed, it is well known that the first few days of life are a crucial time frame for the correct development and modulation of the human gut microbiota (5–7). Therefore, inappropriate seeding of the microbial communities occurring at childbirth due to defects of optimal microbial acquisition, e.g., vertical mother-to-infant transmission, may have a short- or long-term impact on the host health (8–12).

In this regard, it has been observed that an alteration of the taxonomic composition of the gut bacterial community and their functional properties characterize the gut microbiota of NEC patients compared with those of healthy infants (4, 13). More specifically, intestinal microbial communities of healthy breastfed infants are dominated by bifidobacterial species, mainly *Bifidobacterium bifidum* and *Bifidobacterium longum* subsp. *infantis* (14, 15). In contrast, the gut microbiota of NEC patients showed an increased abundance of *Clostridium* and Enterobacteriaceae genera (16, 17). In this context, an exacerbated pro-inflammatory cascade arising from dysfunctional, or overstated, immunological response to high levels of intestinal lipopolysaccharides (LPS) has been proposed as one possible pathway that

predisposes the infant to NEC pathogenesis (18). Along with the above-mentioned direct host-microbe interactions mediated by the immune system, the gut microbiota was shown to exert a broad physiological effect on the host biochemistry through the gut microbiota metabolome, i.e., microbial-derived secondary metabolites. From this perspective, it becomes evident that perturbation of the microbial community composition may modify the intestinal metabolic profiling, with a subsequent impact on the host's health. Overall, the proven importance of perinatal microbial exposures in health and illness provides the foundation for assuming that bacteria colonizing the infant gut in the immediate post-natal period may be involved in NEC development(19). Nevertheless, identifying specific causative microorganisms, known as microbial biomarkers, remains elusive (20).

In this study, we performed a metagenomics analysis of 171 preterm infant fecal samples, aiming to assess the infant gut microbiota composition during NEC events compared to those of gestational age-matched healthy infants. Furthermore, to identify putative microbial biomarkers of NEC, the obtained metagenomic datasets were also employed to determine the metabolic reactions profiles and the presence of potentially damaging microbial metabolites enriched in NEC.

RESULTS AND DISCUSSION

General features of datasets included in the meta-analysis

A collection of 124 shotgun metagenomic datasets from four different studies (27–31) was retrieved from publicly available repositories (Table S1). Out of these, 67 corresponded to fecal samples of preterm infants suffering from NEC, while the remaining 57 samples were acquired from premature infants considered overall healthy. More precisely, among the NEC subjects, 53 were newborns with confirmed NEC at the time of sampling, while 14 fecal samples were collected prior to NEC diagnosis (Table S1). Gestational age ranged from 23 to 39 weeks, corresponding to a birth weight between 900 and 3010 g. Notably, only four infants showed a gestational age or birth weight longer than 32 weeks and 1500g, respectively (Table S1). As gestational age, birth weight, and post-natal age are among the main factors strongly affecting the infant gut microbiota composition, the selection of only infants born very prematurely with low birth weight and obtaining fecal samples from similar post-natal age enabled us to establish a reliable approach for the comparison of the microbiota composition between these samples. However, according to what was previously shown, the gut microbiota seems to undergo only minor changes up to 6 months of age (32).

Additionally, 47 fecal samples of preterm infants from NICU at Croix rousse university Hospital were also included in the analyses (Table S2). These latter were collected weekly during the first 30 days of life of 18 infants born between 25 and 30 weeks of gestation. While 11 infants did not display any intestinal morbidity (for a total of 24 fecal samples), seven infants developed NEC, developing in total 12 fecal samples before and 11 fecal samples after NEC development and diagnosis (Table S2).

Overall, a total of 171 fecal samples of preterm infants, encompassing 64 cases of ongoing NEC, 26 patients that later established NEC but did not show symptoms at the time of collection, and 81 healthy control samples were evaluated, representing one of the largest shotgun metagenomics data collections to date. All the datasets were re-analyzed using the same analysis pipeline, i.e., METAnnotatorX (33).

Gut microbiota variability between cases of NEC and healthy subjects

In order to highlight the differences in gut microbiota composition between infants with manifested NEC symptoms and healthy subjects, we compared the microbiota composition of the retrieved 64 NEC samples with the 81 healthy samples, while the 26 pre-NEC samples were investigated separately (see below). As generally expected for premature newborns, the index of bacterial species richness, i.e., biodiversity, calculated as the number of taxa with a relative abundance of sequenced reads greater than 0.5 %, was on average relatively lower compared to term infants (34, 35), but still statistically higher in the healthy subjects (8.4 ± 5.7) compared to those affected by NEC (6.6 ± 3.7, t-test p-value < 0.05) (detailed data are reported in Table S3).

This statistical reduction of the gut microbial biodiversity associated with NEC is supported by the observation that the two most abundant microbial species in the fecal samples of NEC patients cover 38 % of the whole bacterial population compared to 29.76 % of the healthy infants. Thus, suggesting that even the loss of few species may markedly disrupt the delicate ecological equilibrium established in the very early stages of life in the human gut environment.

The analysis of inter-sample variability of the gut microbiota composition revealed significant differences between infants with NEC diagnosis and control samples (PERMANOVA p-value < 0.05) (Figure S1), reflecting the notion that the gut microbiome of newborns with NEC not only displayed lower biodiversity but also different taxonomic composition from that of their healthy counterparts (36).

For each metagenomic sample, shotgun methods allowed the taxonomic classification of bacterial taxa at the species level, with the estimation of their relative abundance (expressed as percentages of total sequenced reads per sample). Statistical identification of bacterial species with a relative differential abundance between the 81 control and the 64 overt cases of NEC revealed that *Escherichia coli,* and *Enterococcus faecalis* were the main taxa with statistically higher relative abundance in fecal samples of NEC patients, with an average abundance of 26.27 % ± 41.06 % and 11.24 % ± 26.08 % respect to 13.85 % ± 29.14 % and 1.45 % ± 4.06 % in control samples, respectively (ANOVA p-value < 0.01) (Table S3). In contrast, *Streptococcus agalactiae* was the dominant taxon in premature control subjects (average abundance of 15.91 % ± 35.03 % in comparison to 0.28 % ± 1.77 % within NEC population, ANOVA p-value < 0.001) (Table S3), while both groups showed comparable levels of *Staphylococcus epidermidis* and *Klebsiella pneumoniae* indicating at first glance that these bacterial taxa may not be directly related to NEC onset (Table S3).

However, the -microbial profiles of the 145 collected infant gut metagenomic samples have markedly shown a high inter-individual variability, suggesting that infant-specific factors and the different NICU environment may considerably impact the initial microbial colonization of the infants' gut after birth, as previously reported (Table S3) (37, 38).

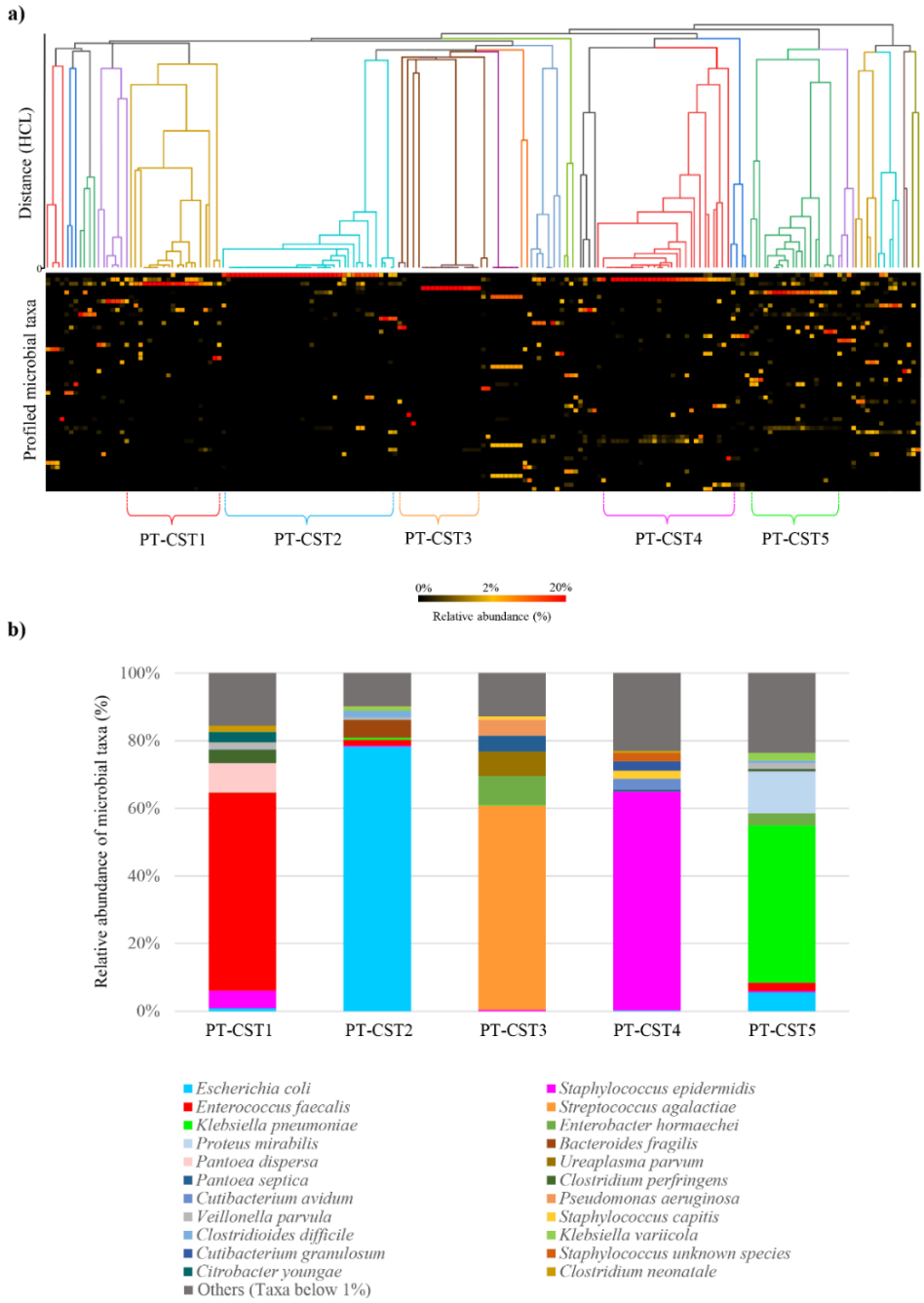


Figure 1. Identification of the five PT-CSTs. Panel **(a)** shows a cladogram of the 171 preterm infant fecal samples obtained through Hierarchical Clustering (HCL) analysis. The cladogram highlighted the five PT-CSTs identified through HCL analysis. Below is reported an overview of the taxonomic composition of the infant population. Panel **(b)** displays the average relative abundance of microbial species of the identified PT-CST, with relative abundances on the vertical axis and the sample on the horizontal axis.

Gut Community State Types (CSTs) distribution in preterm infants.

The high taxonomic variability of the preterm gut microbial communities dramatically reduces the ability to detect significant microbial biomarkers. For this reason, we performed Pearson index-based Hierarchical Clustering Analysis (HCA) employing species-level microbial profiling data to elucidate the prevalence of specific taxonomic patterns, also known as Community State Types (CSTs), across the collected samples (Table S4, Figure 1a). The statistically validated clusters that encompassed at least ten samples have been taken into account as the most prevalent representatives of preterm infants (39) (Table S4, Figure 1a). Accordingly, we identified five archetypical subgroups named Pre-Term Community State Types (PT-CST1-5) (Table S4, Figure 1b), not correlated to specific geographical origins (Table S1). A detailed description of each PT-CST is provided in the Supplementary Text.

Notably, the five PT-CSTs were unevenly distributed across the fecal samples of healthy control and NEC preterm infants, thus revealing the absence of clear associations between specific taxonomic patterns and the development of NEC (Table S4). Nevertheless, within each PT-CST (except for PT-CST3), a statistically significant reduction in biodiversity was observed in NEC rather than in control infants (species-richness average ranged from 4.1 to 8.6 in NEC infants vs. 7 to 19 in control, t-test p-values < 0.01), accompanied by PT-CST-specific increase in relative abundances of well-known (opportunistic) pathogens, such as *E. faecalis*, *E. coli*, *St. epidermidis*, *Clostridioides difficile*, *Ureaplasma parvum*, *Pseudomonas aeruginosa*, *Pseudomonas nosocomialis* and members of the *Klebsiella* genus (Table S5, Figure 2). Notably, most of the latter species have been reported to exert pathogenic outcomes proportionally to their abundance (40–47). These data suggest that the loss of specific taxa may play a direct role in the overgrowth of (opportunistic) pathogens supporting NEC development.

Detailed screening within each PT-CST for taxa involved in the loss of biodiversity revealed that a range of protective/health-promoting appears reduced or absent in NEC subpopulations (Table S5, Figure 2), including members of the genus *Bifidobacterium* (such as *B. bifidum* and *Bifidobacterium breve*) and *Akkermansia* (48–50). In addition, common early infant gut commensals, such as species of the genera *Actinomyces*, *Schaalia*, *Veillonella*, *Bacteroides,* and *Streptococcus,* appeared to be reduced or absent in the fecal samples of NEC patients(39, 51). Notably, since these latter are known to be among the early gut commensals, they may participate in the homeostasis of gut bacterial communities in preterm infants by acting as neutral commensals. Thus, their early loss, beyond cause the overall drop in biodiversity, probably support the overgrowth of the taxa involved in NEC pathogenesis.

Figure 2. Statistically significant differences in the taxonomic composition between NEC and healthy samples of each PT-CST. In detail, for each PT-CST, significant *p*-values obtained by comparison between microbial average abundances of healthy and NEC samples have been highlighted in green. Additional taxa above 1 % of average abundance have been reported.

Gut bacterial network community structure

To assess the ecological role of the low abundance taxa in the preterm infant gut, we explored the microbial communities structure using co-occurrence network analysis (Table S6). Modularity analysis revealed the presence of 13 clusters of co-variating species, which were highlighted with different node colors (Figure 3). As depicted in Figure 3, the five microbial species predominant in the PT-CSTs1-5 correspond to the higher interconnected nodes, suggesting that they are also more linked with other infant gut microbiota members. Furthermore, most of the analyzed bacterial taxa are engaged in widespread negative relationships, while the positive ones were observed among minority members of the preterm infant gut, which give rise to a web of interactions supporting the growth of the predominant taxa (Figure 3) (52).

For example, dominant species of PT-CSTs such as *S. agalactiae* (node 156)*, E. coli* (node 151)*, E. faecalis* (node 150), and *K. pneumoniae* (node 148) were engaged in mutual relationships with known low-abundance members of the preterm infant gut microbiota, including *Streptococcus*, *Cutibacterium*, *Enterobacter*, and *Corynebacterium* genera (green group), as well as members of the Actinobacteria family (violet group) (Figure 3). This finding suggests that these specific minor players of the bacterial population can markedly shape the infant gut microbiota community by playing a potentially crucial role in sustaining the balance of the interaction network. Thus, a loss of marginal species could drive the disruption of the intricate microbial network, allowing for potential persistent colonization by (opportunistic) pathogens. Overall, these network analyses provide an overview of the bacterial interactions underlying the predominance of the taxa observed in PT-CSTs, adding new dimensions to our understanding of gut dysbiosis in neonates affected by NEC.

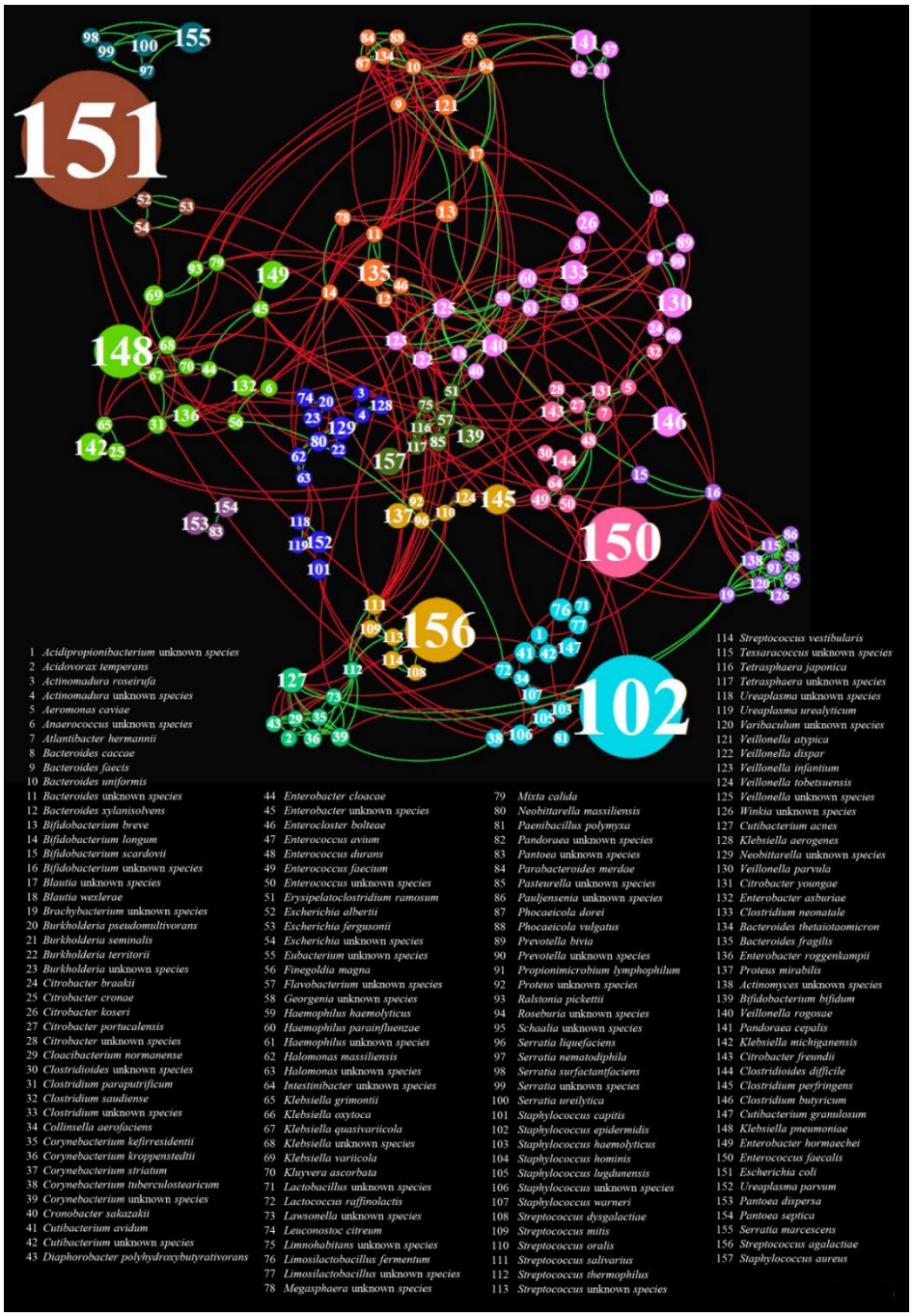


Figure 3. Co-variance of the most abundant microbial species of the preterm infant gut. The force-driven network was constructed using bacterial taxa as nodes and covariances as edges. Red edges correspond to negative correlations, while green edges represent positive associations. The node size is proportional to the degree of interactions, while node colors represent the 13 obtained clusters of co-variating species.

Assessing of gut microbial metabolic pathways in NEC and healthy subjects.

As mentioned above, the gastrointestinal microbial inhabitants may influence host health through their metabolic activities, which participate in bio-modification or *de novo* synthesis of metabolites. Therefore, changes in the microbiome gene repertoire, reflecting shifts in the gut microbiota composition, were analyzed to understand how potential gut bacterial community-derived metabolites differ between NEC patients and healthy subjects, with peculiar focus on essential reactions of metabolic pathways, as reported by literature and MetaCyc database (53) (Table S7). Moreover, potential association between gut-associated bacterial communities and the microbial metabolic pathways was assessed by a Pearson correlation analysis. Specifically, a Pearson coefficient was calculated for each species and each microbial enzyme resulting significantly increased or decreased in NEC subjects compared to healthy infants (Table S8).

In particular, in healthy infants, we detected enzymes related to glycosylated proteins degradation, i.e., α-fucosidase (EC 3.2.1.51) and sialidase (EC 3.2.1.18) that resulted almost entirely absent in NEC microbiomes (Table S7, Figure 4). These enzymes are essential for releasing L-fucose and sialic acid from host-derived glycans such as intestinal mucins and human milk oligosaccharides (HMOs). Indeed, they are typically associated with gut commensals with health-promoting properties which strictly co-evolved with the human host, such as *Bifidobacterium longum*, *Bifidobacterium breve*, and *Bifidobacterium bifidum* (52, 54–56). Moreover, the covariance analysis highlighted that α-fucosidase and sialidase were positively associated with members of *Blautia*, *Cutibacterium* and *Enterobacter* genus and negatively with *E. coli* (Table S8).

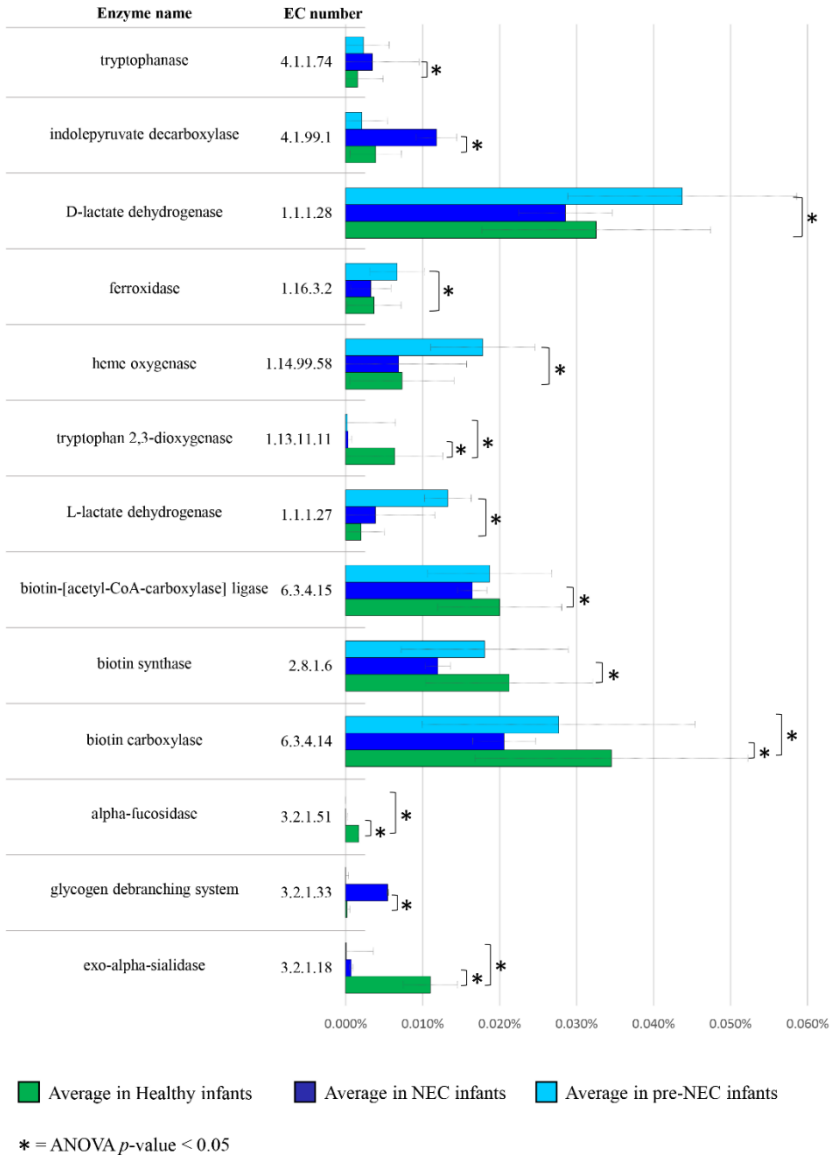


Figure 4. Statistically significant enzymes differential encoded by gut microbiota of healthy, NEC, and pre-NEC infants. Bar plot depicts the relative average abundance of each considered enzyme in healthy, NEC and pre-NEC samples on horizontal axis. The table on the left reports the corresponding EC numbers and enzyme names.

Among the bacterial tryptophan degradation pathways, the key enzymes tryptophanase (TnaA, EC 4.1.99.1) as well as indolepyruvate decarboxylase (EC 4.1.1.74) catalyzing the production of indole and indole 3-acetic acid (IAA) biosynthesis, respectively, showed a 3-fold increase in NEC subjects compared to

healthy infants (ANOVA p-value < 0.05) (Table S7, Figure 4). Notably, as confirmed by positive correlation index reported in Table S8, this may be the result of an higher abundance in NEC individuals of mainly *E. coli*, as well as members of *Klebsiella* and *Staphylococcus* genus, reported to produce these enzymes (57). Although these metabolites are known to have a role in regulating intestinal immunity acting as aryl hydrocarbon receptor (AhR) ligands, their effects are subjected to dietary tryptophan intestinal availability. The broad gastrointestinal damages experienced by NEC infants largely interfere with tryptophan assimilation, likely resulting in tryptophan depletion at the gut level and, therefore, lack of its microbial metabolization.

Although marginally, a portion of tryptophan (TRP) may be metabolized through the kynurenine pathway (KP), whose downstream metabolites, such as kynurenine (Kyn), quinolinic acid, picolinic acid, and kynurenic acid (KA), are known for their neuroactive properties, also regulating various (bio) processes related to inflammation and immune response (58–61). The enzyme tryptophan 2,3-dioxygenase (TDO, EC 1.13.11.11), catalyzing the first and rate-limiting step of the KP, was detected with a 4-fold increase in healthy subjects (ANOVA p-value < 0.05) (Table S7, Figure 4). As previously suggested, although TDO is typically a eukaryotic enzyme, these results supported the notion that some bacteria could synthesize Kyn by expressing an enzyme homologous to TDO (62). Thus, suggesting that depletion of bacterial homologous for this enzyme in NEC subjects may reflect increased gut inflammation and/or unbalanced intestinal mucosal reactivity (63, 64). Remarkably, positive associations were found between TDO abundance and the presence of various commensals of the infant gut, including *Bacteroides*, *Cutibacterium*, and *Enterobacter* genera, while negative correlation were identified with *E. coli* and *S. epidermidis*. Furthermore, biotin (vitamin B7) metabolism-related enzymes, including biotin-(Acetyl-CoA carboxylase) ligase (EC 6.3.4.15), biotin carboxylase (EC 6.3.4.14), and biotin synthase (EC 2.8.1.6) which catalyzes the essential reaction

in the biotin biosynthetic pathway, were found down-represented in NEC patients (decreased respectively by 18.58 %, 40.64 %, 43.85 %, as compared to healthy samples (ANOVA, *p*-value < 0.05) (Table S7, Figure 4). Interestingly, biotin has a putative impact on a range of catabolic and anabolic pathways of the human host, such as carbohydrates and amino acids catabolism or fatty acids synthesis (64), and it can be synthesized at intestinal level by gut commensals possessing the complete biosynthesis pathway, such as *Bacteroides fragilis,* which is a bacterial taxon enriched in healthy infants (66). Furthermore, negative correlation coefficient was observed between component of biotin biosynthetic pathway and *E. coli* (Table S8). In contrast, among the microbial pathways over-expressed in the fecal samples of NEC patients, we found the glycogen debranching enzyme (EC 3.2.1.196) 10-fold higher compared to controls (ANOVA *p*-value < 0.05) (Table S7, Figure 4). This enzyme is essential to complete the glycogen breakdown through the glycogen metabolism pathways(67) and has been referred to as a potential virulence factor in many microorganisms. Indeed, enzymes involved in this complex carbohydrate catabolism might participate in pathogen infectivity, contributing to virulence, colonization, and the environmental survival of pathogen strains. Thus reflecting the increased abundance of *E. coli* and other members of the Gammaproteobacteria class (68) such as *Pseudomonas* and *Klebsiella*, observed in NEC infants compared to healthy subjects (69–71).

Gut microbial community analysis in pre-NEC samples.

In order to identify possible microbial signatures causatively involved in NEC development, we compared the taxonomic profiles of 26 fecal samples collected from preterm infants before NEC development (pre-NEC) versus all the 81 healthy controls. Focusing on the gut bacterial community composition of pre-NEC infants, HCA identified three distinct recurrent taxonomic profiles, named Pre-NEC

Community State Types (PN-CST1-3), that can be observed in preterm infants who later developed NEC (Table S9, Figure 5). Notably, each PN-CST appears to be dominated by one or few recognized (opportunistic) pathogens.

In particular, despite PN-CST1 seems to be dominated by common gut colonizers such as *E. coli*, *E. faecalis,* or members of the *Prevotella* genus, the two most abundant species observed in each profiled sample constitute >90 % of the whole microbial population, thus indicating a marked simplification of the microbial population, i.e., reduced biodiversity (Table S9).

Figure 5. Species-level taxonomic composition of pre-NEC infants. Pre-NEC samples were grouped according to the specific PN-CST to which they belong. Relative abundances are reported on the vertical axes and the samples on the horizontal axis.

In contrast, PN-CST2 and PN-CST3 showed higher taxonomic complexity. Still, they were dominated by well-recognized pathogens frequently involved in nosocomial infections such as *K. pneumoniae* (72) and *Proteus mirabilis* (73, 74) in PN-CST2 or *Clostridium perfringens* (75), *Clostridium neonatale* (76), *Pantoea dispersa* (77, 78) and *Staphylococcus aureus* (79) in PN-CST3 (Table S9, Figure 5). Remarkably, the latter pathogens are absent or present at a much lower relative abundance in healthy preterm infants who did not develop NEC (*K. pneumoniae* and *P. mirabilis* showed an average abundance of 43.49 % and 31.51 % in PN-CST2, respectively, in contrast to 3.83 % and 0 % in control, respectively; *C. perfringens*, *C. neonatale*, *P. dispersa*, and *S. aureus* displayed an average abundance of 26.52 %, 6,85 %, 8.14 % and 5,12 % in PN-CST3, respectively, rather than 1.02 %, 1.21 %, 0 %, and 0.65 % in healthy subjects, respectively, t-test p-values < 0.01).

In contrast, *P. mirabilis*, *C. perfringens*, *C. neonatale,* and *P. dispersa* were found to be almost absent also in NEC-microbiomes (*C. perfringens*, and *C. neonatale* showed an average abundance of 0.62 %; and 0.12 %, respectively, while *P. mirabilis*, and *P. dispersa* were absent), likely as a consequence of the broad-spectrum antibiotic treatments. Thus, suggesting that the above-mentioned taxa could represent early microbial signatures to be further investigated. Nevertheless, culture-based analyses of NEC microbiota should be performed to corroborate the relevance of microbial biomarkers as putative targets for non-invasive screening able to select infants with "high risk" microbial patterns. Notably, this CST-based overview suggested that the development of NEC should be considered a multifactorial event that involves a loss in biodiversity accompanied by the overgrowth of specific opportunistic pathogens based on the taxonomic pattern initially established in each preterm infant.

In order to identify putative novel metabolic biomarkers of NEC that can support early diagnosis of this disease, we performed the prediction of microbial enzymatic

reactions based on the shotgun metagenomics data of the 26 pre-NEC infant fecal samples. Interestingly, among the bacterial metabolites which were significantly increased before clinical development of NEC, we detected enzymes involved in iron uptake and heme degradation, i.e., ferroxidase (EC 1.16.3.2) and heme oxygenase (EC 1.14.99.58) (Table S10, Figure 4), whose activities are essential for Gram-negative pathogens, which obtain the iron requested for their own growth from heme sequestered from their hosts (80, 81). In particular, such microbial enzymes showed an increase of 45 % and 143 % in the pre-NEC stage with respect to healthy controls (Table S10, Figure 4).

Furthermore, the metabolites profiling of gut bacteria also revealed that the enzymes L- and D-lactate dehydrogenase (EC 1.1.1.27 and EC 1.1.1.28, respectively) were more abundant in pre-NEC infants compared to their healthy counterpart showing a 373 % and 34 % enhance, respectively (t-test p-value < 0.05) (Table S10, Figure 5). In contrast, compared to the control group, infants who received NEC diagnosis did not show such a high abundance of DL-lactate dehydrogenase (t-test p-value > 0.05), likely due to the broad antibiotics treatments as well as invasive medical and surgical procedures leading to the disruption of the gut microbiota and hence of the lactate-producing bacteria (82). However, a recent study reported higher levels of lactate in the urine metabolome of preterm newborns with NEC compared to the control group, reinforcing the possible connection between this metabolite and NEC onset (83). Besides, DL-lactate has been reported to accumulate in feces from individuals with ulcerative colitis and inflammatory bowel disease (84, 85), highlighting the potential negative consequences of elevated intestinal lactate levels leading to gut acidosis due to its pH lowering effect (86, 87). Moreover, it has been demonstrated that lactate accumulation may occur due to a decrease in lactate-utilizing bacteria, such as members of Bacteroidetes and Firmicutes phyla (88–90). These observations suggested that an imbalance between gut communities of lactate-producing and

lactate-utilizing bacteria leading to gastrointestinal DL-lactate accumulation can arise in infants who later develop NEC. Thus, it can represent a possible predictive biomarker of NEC although future experimental studies employing analytical methods are needed to validate the clinical value of lactate level in infant stool at pre-NEC stage.

Conclusion

Necrotizing enterocolitis (NEC) is a relatively common severe gastrointestinal disease mainly affecting low birth-weight infants, with a reported mortality rate ranging between 15% and 30% (91). Recently, the increased prevalence of this intestinal disease has been mainly linked with early gut microbial dysbiosis and dysregulated immune response leading to overemphasized gastrointestinal inflammation in preterm infants.

In order to explore the gut microbial communities before and during NEC, we performed one of the largest shotgun metagenomics meta-analysis, including 124 publicly available datasets of preterm infant fecal samples along with 47 shotgun metagenomics data obtained from fecal samples of 18 preterm infants collected across the first month of life in the framework of this study at the NICU Croix rousse university Hospital. Statistically validated hierarchical clustering analysis allowed the identification of five archetypical recurring taxonomic profiles named Pre-Term Community State Types encompassing both healthy controls and NEC-affected preterm infants. Investigation of the NEC sub-population of each PT-CST revealed a common trend of biodiversity loss accompanied by a marked increase in renowned (opportunistic) pathogens. Moreover, reduction in biodiversity reflected the loss of commensal or protective/health-promoting taxa following NEC development, including members of the genera *Bifidobacterium* and *Akkermansia*. The ecological role of minor components of the gut microbial population of preterm infants was

further explored by covariance analysis. The resulting force-driven network revealed that specific clusters of microorganisms, constituted by both dominant and low-abundance taxa, tend to co-variate. Thus, indicating an intricate network of ecological relationships explaining how the loss of minor players is responsible for such broad impacts on the whole taxonomic profile. Overall, this CST-based overview of taxonomic features associated with NEC patients or healthy preterm infants revealed that biodiversity reduction driven by the loss of specific taxa might represent valuable microbial biomarkers for the early diagnosis of NEC and a starting point for the development of novel bacteria-based therapies. However, the clinical significance of these possible translational outcomes will need further validation in following up clinical trials.

The shotgun metagenomics data were also employed for the investigation of the occurrence of metabolic pathways. Intriguingly, we revealed that enzymes involved in tryptophan metabolism, biotin synthesis, and HMOs degradation were depleted in samples of preterm infants diagnosed with NEC compared to healthy controls, while the enzyme lactate dehydrogenase (LDH) was overabundant in preterm infants prior to NEC development. Hence, suggesting a gastrointestinal DL-lactate accumulation arising in infants who later develop NEC, making this compound a potential functional biomarker for early diagnosis of NEC.

MATERIAL AND METHODS

Ethical statement

The human study protocol was approved by the local Ethics Committee (Comité de Protection des Personnes Sud-Est IV, Lyon).

Data collection

To perform a meta-analysis, we retrieved four publicly available datasets from studies involving the taxonomic evaluation of the gut microbiota in preterm infants with and without NEC (PRJNA46337, PRJNA376566, PRJNA63661, and PRJNA273761). Remarkably, we selected shotgun metagenomics datasets achieved by the Illumina sequencing platform to obtain high resolution and limited input data variability. Accordingly, we collected 124 fecal samples from five different Neonatal Intensive Care Unit (NICU) in United Stated (USA), corresponding to 57 infants considered overall healthy and 67 declared affected by NEC.

Additional 47 fecal samples from 18 premature infants enrolled in NICU at Croix rousse university Hospital of Lyon, France, were considered between September 2014 and November 2014. Specifically, fecal samples of infants were collected weekly during their first 40-days NICU stay. Seven infants developed NEC (Stage II and Stage III NEC diagnosis) (21), while 11 with no infectious intestinal complications have been regarded as controls.

Sample collection and DNA extraction

Stool samples were kept on ice immediately after collection and shipped to the laboratory under frozen conditions, where they were preserved at –20°C until they were processed. DNA extraction from each sample was performed using the QIAmp DNA Stool mini-kit following the manufacturer's instructions (Qiagen, Germany). DNA concentration and purity were determined employing a Picodrop microliter Spectrophotometer (Picodrop).

Sequencing and taxonomic classification

The shotgun metagenomic sequencing was performed by GenProbio Srl (www.genprobio.com). DNA library preparation was performed using the Nextera XT DNA sample preparation kit (Illumina, San Diego, CA), following the

manufacturer's instructions. One ng input DNA from each sample was used for library preparation. The isolated DNA underwent fragmentation, adapter ligation, and amplification. The ready-to-go libraries were pooled equimolarly, denatured, and diluted to a sequencing concentration of 1.5 pM. Sequencing was performed on NextSeq 550 instrument (Illumina, San Diego, CA), following the manufacturer's instructions, using the 2x150 bp High Output sequencing kit, and spike-in of 1 % PhiX control library. Shotgun metagenomics analysis of the fecal samples produced an average number of 3,314,112.39 ± 6,556,669.69 reads per sample. The raw data in fastq format, including those retrieved from publicly available shotgun data sets, were submitted to quality filtering for the removal of reads with an average quality < 25. Subsequently, human DNA was removed by reads mapping on the *Homo sapiens* genome, obtaining an average of 13,891.71 ± 3,580.60 reads per sample that were submitted to downstream analyses.

Following, taxonomic profiling of retained reads was performed with the METAnnotatorX bioinformatics platform (22). Remarkably, in order to homogenize the sequencing data from different studies, we selected only those produced through Illumina sequencing method. Taxonomic classification of each sequenced read was achieved using MegaBLAST (23), employing the curated non-redundant sequence database of genomes retrieved from the National Center for Biotechnology Information (NCBI). For each metagenomic sample, species richness, i.e. biodiversity, represented the number of gut-associated bacterial taxa whose sequenced reads had e relative abundance greater than 0.5 %. Similarities between samples (beta-diversity) were calculated by Bray-Curtis dissimilarity based on species abundance. The range of similarities is calculated between values 0 and 1. PCoA representation of beta-diversity was performed using ORIGIN 2021 (https://www.originlab.com/2021). In the PCoA, each dot represented a sample distributed in tridimensional space according to its own bacterial composition.

Functional prediction

Functional profiling of sequenced reads was obtained through the METAnnotatorX bioinformatics platform (24). In addition, enzymatic reactions profiles were predicted using METAnnotatorX based on the MetaCyc database (25).

Microbial co-occurrence and network analyses

Covariance analysis involving the 233 bacterial species obtained by shotgun profiling of the 171 infant fecal samples was accomplished employing Kendall's tau rank covariance analysis (26). Using software Gephi (https://gephi.org/), such correlation coefficients were then exploited to build a force-driven network, where bacterial species, represented in the form of nodes, are connected by edges. Each node size is proportional to the number of interactions of a specific taxon, i.e., the node degree, while the edge color indicates the type of interaction, i.e., positive (green) or negative (red).

Statistical analyses

Sample clustering based on different predominant taxa was achieved by Hierarchical Clustering Analyses (HCL) using bacterial abundance information at the species level and was then calculated through TMeV 4.8.1 software using Pearson correlation as a distance metric. The data obtained were represented in the form of a cladogram. ORIGIN 2021 (https://www.originlab.com/2021) and the online version of Medcalc software (https://www.medcalc.org/) were used to compute statistical analyses, including t-test, ANOVA test, and Chi-squared test. PERMANOVA analyses were performed using 1000 permutations to estimate *p*-values for differences among populations in PCoA analyses. Furthermore, the differential abundance of bacterial genera was tested by t-test analysis. Covariance analyses between microbial community and bacterial metabolic pathways was obtained through a Pearson

correlation analysis between the abundances of the species and abundances of each individual enzyme observed in the profiled datasets.

Data Availability Statement

Shotgun metagenomics data have been deposited in the NCBI Short Read Archive (SRA) under BioProject code PRJNA733860.

ACKNOWLEDGMENTS

We thank GenProbio Srl for the financial support of the Laboratory of Probiogenomics. Part of this research is conducted using the High Performance Computing (HPC) facility of the University of Parma.

References

1. Zozaya C, García González I, Avila-Alvarez A, Oikonomopoulou N, Sánchez Tamayo T, Salguero E, Saenz de Pipaón M, García-Muñoz Rodrigo F, Couce ML. 2020. Incidence, Treatment, and Outcome Trends of Necrotizing Enterocolitis in Preterm Infants: A Multicenter Cohort Study. Frontiers in Pediatrics 8.

2. Neu J, Walker WA. 2011. Necrotizing Enterocolitis. New England Journal of Medicine 364:255–264.

3. Claud EC, Walker WA. 2001. Hypothesis: inappropriate colonization of the premature intestine can cause neonatal necrotizing enterocolitis. The FASEB Journal 15:1398–1403.

4. Brehin C, Dubois D, Dicky O, Breinig S, Oswald E, Serino M. 2020. Evolution of Gut Microbiome and Metabolome in Suspected Necrotizing Enterocolitis: A Case-Control Study. Journal of Clinical Medicine 9:2278.

5. Milani C, Duranti S, Bottacini F, Casey E, Turroni F, Mahony J, Belzer C, Delgado Palacio S, Arboleya Montes S, Mancabelli L, Lugli GA, Rodriguez JM, Bode L, de Vos W, Gueimonde M, Margolles A, van Sinderen D, Ventura M. 2017. The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. Microbiology and Molecular Biology Reviews 81.

6. Saturio S, Nogacka AM, Suárez M, Fernández N, Mantecón L, Mancabelli L, Milani C, Ventura M, de los Reyes-Gavilán CG, Solís G, Arboleya S, Gueimonde M. 2021. Early-life development of the bifidobacterial community in the infant gut. International Journal of Molecular Sciences 22.

7. Milani C, Duranti S, Bottacini F, Casey E, Turroni F, Mahony J, Belzer C, Delgado Palacio S, Arboleya Montes S, Mancabelli L, Lugli GA, Rodriguez JM, Bode L, de Vos W, Gueimonde M,

Margolles A, van Sinderen D, Ventura M. 2017. The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. Microbiology and Molecular Biology Reviews 81.

8. Rautava S, Luoto R, Salminen S, Isolauri E. 2012. Microbial contact during pregnancy, intestinal colonization and human disease. Nature Reviews Gastroenterology and Hepatology. Nat Rev Gastroenterol Hepatol.

9. Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, Knight R. 2010. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proceedings of the National Academy of Sciences of the United States of America 107:11971–11975.

10. Forsgren M, Isolauri E, Salminen S, Rautava S. 2017. Late preterm birth has direct and indirect effects on infant gut microbiota development during the first six months of life. Acta Paediatrica, International Journal of Paediatrics 106:1103–1109.

11. Wang S, Ryan CA, Boyaval P, Dempsey EM, Ross RP, Stanton C. 2020. Maternal Vertical Transmission Affecting Early-life Microbiota Development. Trends in Microbiology. Elsevier Ltd.

12. Kundu P, Blacher E, Elinav E, Pettersson S. 2017. Our Gut Microbiome: The Evolving Inner Self. Cell. Cell Press.

13. Carlisle EM, Morowitz MJ. 2013. The intestinal microbiome and necrotizing enterocolitis. Current Opinion in Pediatrics. Curr Opin Pediatr.

14. Yang B, Chen Y, Stanton C, Ross RP, Lee YK, Zhao J, Zhang H, Chen W. 2019. Bifidobacterium and lactobacillus composition at species level and gut microbiota diversity in infants before 6 weeks. International Journal of Molecular Sciences 20.

15. Saturio S, Nogacka AM, Suárez M, Fernández N, Mantecón L, Mancabelli L, Milani C, Ventura M, de los Reyes-Gavilán CG, Solís G, Arboleya S, Gueimonde M. 2021. Early-life development of the bifidobacterial community in the infant gut. International Journal of Molecular Sciences 22.

16. Pammi M, Cope J, Tarr PI, Warner BB, Morrow AL, Mai V, Gregory KE, Simon Kroll J, McMurtry V, Ferris MJ, Engstrand L, Lilja HE, Hollister EB, Versalovic J, Neu J. 2017. Intestinal dysbiosis in preterm infants preceding necrotizing enterocolitis: A systematic review and meta-analysis. Microbiome 5.

17. Wandro S, Osborne S, Enriquez C, Bixby C, Arrieta A, Whiteson K. 2018. The Microbiome and Metabolome of Preterm Infant Stool Are Personalized and Not Driven by Health Outcomes, Including Necrotizing Enterocolitis and Late-Onset Sepsis. mSphere 3.

18. Niemarkt HJ, de Meij TG, van Ganzewinkel CJ, de Boer NKH, Andriessen P, Hütten MC, Kramer BW. 2019. Necrotizing Enterocolitis, Gut Microbiota, and Brain Development: Role of the Brain-Gut Axis. Neonatology 115:423–431.

19. Shokunbi MT, Gelb AW, Wu XM, Miller DJ. 1990. Continuous lidocaine infusion and focal feline cerebral ischemia. Stroke 21:107–111.

20. Coggins SA, Wynn JL, Weitkamp JH. 2015. Infectious causes of necrotizing enterocolitis. Clinics in Perinatology. W.B. Saunders.

21. Bell MJ, Ternberg JL, Feigin RD, Keating JP, Marshall R, Barton L, Brotherton T. 1978. Neonatal necrotizing enterocolitis. Therapeutic decisions based upon clinical staging. Annals of surgery 187:1–7.

22. Milani C, Casey E, Lugli GA, Moore R, Kaczorowska J, Feehily C, Mangifesta M, Mancabelli L, Duranti S, Turroni F, Bottacini F, Mahony J, Cotter PD, McAuliffe FM, van Sinderen D, Ventura M. 2018. Tracing mother-infant transmission of bacteriophages by means of a novel analytical tool for shotgun metagenomic datasets: METAnnotatorX. Microbiome 6.

23. Chen Y, Ye W, Zhang Y, Xu Y. 2015. High speed BLASTN: An accelerated MegaBLAST search tool. Nucleic Acids Research 43:7762–7768.

24. Milani C, Casey E, Lugli GA, Moore R, Kaczorowska J, Feehily C, Mangifesta M, Mancabelli L, Duranti S, Turroni F, Bottacini F, Mahony J, Cotter PD, McAuliffe FM, van Sinderen D, Ventura M. 2018. Tracing mother-infant transmission of bacteriophages by means of a novel analytical tool for shotgun metagenomic datasets: METAnnotatorX. Microbiome 6.

25. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Karp PD. 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Research 44:D471–D480.

26. Liu X, Ning J, Cheng Y, Huang X, Li R. 2019. A flexible and robust method for assessing conditional association and conditional concordance. Statistics in Medicine 38:3656–3668.

27. Warner BB, Deych E, Zhou Y, Hall-Moore C, Weinstock GM, Sodergren E, Shaikh N, Hoffmann JA, Linneman LA, Hamvas A, Khanna G, Rouggly-Nickless LC, Ndao IM, Shands BA, Escobedo M, Sullivan JE, Radmacher PG, Shannon WD, Tarr PI. 2016. Gut bacteria dysbiosis and necrotising enterocolitis in very low birthweight infants: A prospective case-control study. The Lancet 387:1928–1936.

28. Raveh-Sadka T, Thomas BC, Singh A, Firek B, Brooks B, Castelle CJ, Sharon I, Baker R, Good M, Morowitz MJ, Banfield JF. 2015. Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. eLife 2015.

29. Brooks B, Olm MR, Firek BA, Baker R, Thomas BC, Morowitz MJ, Banfield JF. 2017. Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. Nature Communications 8.

30. Brown CT, Xiong W, Olm MR, Thomas BC, Baker R, Firek B, Morowitz MJ, Hettich RL, Banfield JF. 2018. Hospitalized premature infants are colonized by related bacterial strains with distinct proteomic profiles. mBio 9.

31. Ward D v., Scholz M, Zolfo M, Taft DH, Schibler KR, Tett A, Segata N, Morrow AL. 2016. Metagenomic Sequencing with Strain-Level Resolution Implicates Uropathogenic E. coli in Necrotizing Enterocolitis and Mortality in Preterm Infants. Cell Reports 14:2912–2924.

32. L M, C T, C M, GA L, F F, F T, D van S, M V. 2020. Multi-population cohort meta-analysis of human intestinal microbiota in early life reveals the existence of infant community state types (ICSTs). Computational and structural biotechnology journal 18:2480–2493.

33. Milani C, Casey E, Lugli GA, Moore R, Kaczorowska J, Feehily C, Mangifesta M, Mancabelli L, Duranti S, Turroni F, Bottacini F, Mahony J, Cotter PD, McAuliffe FM, van Sinderen D, Ventura M. 2018. Tracing mother-infant transmission of bacteriophages by means of a novel analytical tool for shotgun metagenomic datasets: METAnnotatorX. Microbiome 6.

34. Chernikova DA, Madan JC, Housman ML, Zain-ul-abideen M, Lundgren SN, Morrison HG, Sogin ML, Williams SM, Moore JH, Karagas MR, Hoen AG. 2018. The premature infant gut microbiome during the first 6 weeks of life differs based on gestational maturity at birth. Pediatric Research 84:71–79.

35. Rougé C, Goldenberg O, Ferraris L, Berger B, Rochat F, Legrand A, Göbel UB, Vodovar M, Voyer M, Rozé JC, Darmaun D, Piloquet H, Butel MJ, de La Cochetière MF. 2010. Investigation of the intestinal microbiota in preterm infants using different methods. Anaerobe 16:362–370.

36. Morrow AL, Lagomarcino AJ, Schibler KR, Taft DH, Yu Z, Wang B, Altaye M, Wagner M, Gevers D, Ward D v., Kennedy MA, Huttenhower C, Newburg DS. 2013. Early microbial and metabolomic signatures predict later onset of necrotizing enterocolitis in preterm infants. Microbiome 1.

37. Henderickx JGE, Zwittink RD, van Lingen RA, Knol J, Belzer C. 2019. The preterm gut microbiota: An inconspicuous challenge in nutritional neonatal care. Frontiers in Cellular and Infection Microbiology. Frontiers Media S.A.

38. Barrett E, Kerr C, Murphy K, O'Sullivan O, Ryan CA, Dempsey EM, Murphy BP, O'Toole PW, Cotter PD, Fitzgerald GF, Ross RP, Stanton C. 2013. The individual-specific and diverse nature of the preterm infant microbiota. Archives of Disease in Childhood: Fetal and Neonatal Edition 98.

39. Mancabelli L, Tarracchini C, Milani C, Lugli GA, Fontana F, Turroni F, van Sinderen D, Ventura M. 2020. Multi-population cohort meta-analysis of human intestinal microbiota in early life reveals the existence of infant community state types (ICSTs). Computational and Structural Biotechnology Journal 18:2480–2493.

40. Delaplain PT, Bell BA, Wang J, Isani M, Zhang E, Gayer CP, Grishin A v., Ford HR. 2019. Effects of artificially introduced Enterococcus faecalis strains in experimental necrotizing enterocolitis. PLoS ONE 14.

41. Stoll BJ, Hansen NI, Sánchez PJ, Faix RG, Poindexter BB, van Meurs KP, Bizzarro MJ, Goldberg RN, Frantz ID, Hale EC, Shankaran S, Kennedy K, Carlo WA, Watterberg KL, Bell EF, Walsh MC, Schibler K, Laptook AR, Shane AL, Schrag SJ, Das A, Higgins RD. 2011. Early onset neonatal sepsis: The burden of group B streptococcal and E. coli disease continues. Pediatrics 127:817–826.

42. Brescó MS, Harris LG, Thompson K, Stanic B, Morgenstern M, O'Mahony L, Richards RG, Moriarty TF. 2017. Pathogenic mechanisms and host interactions in Staphylococcus epidermidis device-related infection. Frontiers in Microbiology. Frontiers Media S.A.

43. Smith AB, Ocana JS, Zackular JP. 2020. From nursery to nursing home: Emerging concepts in clostridioides difficile pathogenesis. Infection and Immunity 88.

44. Glaser K, Gradzka-Luczewska A, Szymankiewicz-Breborowicz M, Kawczynska-Leda N, Henrich B, Waaga-Gasser AM, Speer CP. 2019. Perinatal ureaplasma exposure is associated with increased risk of late onset sepsis and imbalanced inflammation in preterm infants and may add to lung injury. Frontiers in Cellular and Infection Microbiology 9.

45. Xu W, He L, Liu C, Rong J, Shi Y, Song W, Zhang T, Wang L. 2015. The effect of infection control nurses on the occurrence of Pseudomonas aeruginosa healthcare-acquired infection and multidrug-resistant strains in critically-Ill children. PLoS ONE 10.

46. Zhen X, Lundborg CS, Sun X, Hu X, Dong H. 2020. Clinical and economic impact of third-generation cephalosporin-resistant infection or colonization caused by Escherichia coli and Klebsiella pneumoniae: A multicenter study in China. International Journal of Environmental Research and Public Health 17:1–12.

47. Li S, Liu J, Chen F, Cai K, Tan J, Xie W, Qian R, Liu X, Zhang W, Du H, Liu Y, Huang L. 2020. A risk score based on pediatric sequential organ failure assessment predicts 90-day mortality in children with Klebsiella pneumoniae bloodstream infection. BMC Infectious Diseases 20.

48. Liu X, Zhao F, Liu H, Xie Y, Zhao D, Li C. 2021. Transcriptomics and metabolomics reveal the adaption of Akkermansia muciniphila to high mucin by regulating energy homeostasis. Scientific Reports 11:9073.

49. Hagen PC, Skelley JW. 2019. Efficacy of bifidobacterium species in prevention of necrotizing enterocolitis in very-low birth weight infants. A systematic review. Journal of Pediatric Pharmacology and Therapeutics. Pediatric Pharmacy Advocacy Group, Inc.

50. de Andrés J, Manzano S, García C, Rodríguez JM, Espinosa-Martos I, Jiménez E. 2018. Modulatory effect of three probiotic strains on infants' gut microbial composition and immunological parameters on a placebo-controlled, double-blind, randomised study. Beneficial Microbes 9:573–584.

51. Hill CJ, Lynch DB, Murphy K, Ulaszewska M, Jeffery IB, O'Shea CA, Watkins C, Dempsey E, Mattivi F, Tuohy K, Paul Ross R, Anthony Ryan C, O' Toole PW, Stanton C. 2017. Evolution of gut microbiota composition from birth to 24 weeks in the INFANTMET Cohort. Microbiome 5.

52. Turroni F, Milani C, Duranti S, Mahony J, van Sinderen D, Ventura M. 2018. Glycan Utilization and Cross-Feeding Activities by Bifidobacteria. Trends in Microbiology. Elsevier Ltd.

53. R C, R B, CA F, IM K, A K, M K, M L, PE M, Q O, WK O, S P, P S, PD K. 2018. The MetaCyc database of metabolic pathways and enzymes. Nucleic acids research 46:D633–D639.

54. Tailford LE, Crost EH, Kavanaugh D, Juge N. 2015. Mucin glycan foraging in the human gut microbiome. Frontiers in Genetics 5.

55. Egan M, Motherway MOC, Ventura M, van Sinderen D. 2014. Metabolism of sialic acid by Bifidobacterium breve UCC2003. Applied and Environmental Microbiology 80:4414–4426.

56. Bunesova V, Lacroix C, Schwab C. 2016. Fucosyllactose and L-fucose utilization of infant Bifidobacterium longum and Bifidobacterium kashiwanohense. BMC Microbiology 16:1–12.

57. Kaur H, Bose C, Mande SS. 2019. Tryptophan Metabolism by Gut Microbiome and Gut-Brain-Axis: An in silico Analysis. Frontiers in Neuroscience 13.

58. Gao J, Xu K, Liu H, Liu G, Bai M, Peng C, Li T, Yin Y. 2018. Impact of the gut microbiota on intestinal immunity mediated by tryptophan metabolism. Frontiers in Cellular and Infection Microbiology. Frontiers Media S.A.

59. Zelante T, Iannitti RG, Cunha C, DeLuca A, Giovannini G, Pieraccini G, Zecchi R, D'Angelo C, Massi-Benedetti C, Fallarino F, Carvalho A, Puccetti P, Romani L. 2013. Tryptophan catabolites from microbiota engage aryl hydrocarbon receptor and balance mucosal reactivity via interleukin-22. Immunity 39:372–385.

60. Kennedy PJ, Cryan JF, Dinan TG, Clarke G. 2017. Kynurenine pathway metabolism and the microbiota-gut-brain axis. Neuropharmacology. Elsevier Ltd.

61. Jovanovic F, Candido KD, Knezevic NN. 2020. The role of the kynurenine signaling pathway in different chronic pain conditions and potential use of therapeutic agents. International Journal of Molecular Sciences. MDPI AG.

62. Agus A, Planchais J, Sokol H. 2018. Gut Microbiota Regulation of Tryptophan Metabolism in Health and Disease. Cell Host and Microbe. Cell Press.

63. Zelante T, Iannitti RG, Cunha C, DeLuca A, Giovannini G, Pieraccini G, Zecchi R, D'Angelo C, Massi-Benedetti C, Fallarino F, Carvalho A, Puccetti P, Romani L. 2013. Tryptophan catabolites from microbiota engage aryl hydrocarbon receptor and balance mucosal reactivity via interleukin-22. Immunity 39:372–385.

64. Spichak S, Bastiaanssen TFS, Berding K, Vlckova K, Clarke G, Dinan TG, Cryan JF. 2021. Mining microbes for mental health: Determining the role of microbial metabolic pathways in human brain health and disease. Neuroscience and Biobehavioral Reviews. Elsevier Ltd.

65. Biochemistry, Water Soluble Vitamins - PubMed.

66. K Y, K H, K S, J K. 2019. Metabolism of Dietary and Microbial Vitamin B Family in the Regulation of Host Immunity. Frontiers in nutrition 6.

67. Berg JM, Tymoczko JL, Stryer L. 2002. Glycogen Breakdown Requires the Interplay of Several Enzymes.

68. T K, S S, Y U, N F, K U, T S, A K, J A, C C, S B, Y N. 2016. Comparison of Chain-Length Preferences and Glucan Specificities of Isoamylase-Type α-Glucan Debranching Enzymes from Rice, Cyanobacteria, and Bacteria. PloS one 11.

69. Shelburne SA, Davenport MT, Keith DB, Musser JM. 2008. The role of complex carbohydrate catabolism in the pathogenesis of invasive streptococci. Trends in Microbiology. Trends Microbiol.

70. Jones SA, Jorgensen M, Chowdhury FZ, Rodgers R, Hartline J, Leatham MP, Struve C, Krogfelt KA, Cohen PS, Conway T. 2008. Glycogen and maltose utilization by Escherichia coli O157:H7 in the mouse intestine. Infection and Immunity 76:2531–2540.

71. Sekar K, Linker SM, Nguyen J, Grünhagen A, Stocker R, Sauer U. 2020. Bacterial glycogen provides short-term benefits in changing environments. Applied and Environmental Microbiology 86.

72. Jones RN. 2010. Microbial etiologies of Hospital-acquired bacterial pneumonia and ventilator-associated bacterial pneumoniaClinical Infectious Diseases. Clin Infect Dis.

73. Schaffer JN, Pearson MM. 2015. Proteus mirabilis and Urinary Tract Infections. Microbiology Spectrum 3.

74. Armbruster CE, Mobley HLT, Pearson MM. 2018. Pathogenesis of Proteus mirabilis Infection. EcoSal Plus 8.

75. Mehdizadeh Gohari I, A. Navarro M, Li J, Shrestha A, Uzal F, A. McClane B. 2021. Pathogenicity and virulence of Clostridium perfringens . Virulence 12:723–753.

76. Hosny M, Baptiste E, Levasseur A, la Scola B. 2019. Molecular epidemiology of Clostridium neonatale and its relationship with the occurrence of necrotizing enterocolitis in preterm neonates. New Microbes and New Infections 32.

77. Panditrao M, Panditrao M. 2018. Pantoea dispersa: Is it the Next Emerging “Monster” in our Intensive Care Units? A Case Report and Review of Literature. Anesthesia, essays and researches 12:963–966.

78. Mani S, Nair J. 2021. Pantoea Infections in the Neonatal Intensive Care Unit. Cureus 13.

79. Cheung GYC, Bae JS, Otto M. 2021. Pathogenicity and virulence of Staphylococcus aureus. Virulence. Bellwether Publishing, Ltd.

80. Palmer LD, Skaar EP. 2016. Transition Metals and Virulence in Bacteria. Annual Review of Genetics. Annual Reviews Inc.

81. Sheldon JR, Laakso HA, Heinrichs DE. 2016. Iron Acquisition Strategies of Bacterial Pathogens. Microbiology Spectrum 4.

82. AT R, EH C, B K, SP N, NA W, MM V, HK D, S J, FS M, SN N, TM O, JP W, MA D, LA D. 2018. Antibiotic-induced changes in the microbiota disrupt redox dynamics in the gut. eLife 7.

83. JC P, A DM, M M, S C, A D, A N, V F, F CM. 2021. Urine NMR Metabolomics Profile of Preterm Infants With Necrotizing Enterocolitis Over the First Two Months of Life: A Pilot Longitudinal Case-Control Study. Frontiers in molecular biosciences 8.

84. Vernia P, Caprilli R, Latella G, Barbetti F, Magliocca FM, Cittadini M. 1988. Fecal Lactate and Ulcerative Colitis. Gastroenterology 95:1564–1568.

85. Hove H, Nordgaard-Andersen I, Mortensen PB. 1994. Faecal DL-lactate concentration in 100 gastrointestinal patients. Scandinavian Journal of Gastroenterology 29:255–259.

86. Belenguer A, Duncan SH, Holtrop G, Anderson SE, Lobley GE, Flint HJ. 2007. Impact of pH on lactate formation and utilization by human fecal microbial communities. Applied and Environmental Microbiology 73:6526–6533.

87. Wang SP, Rubio LA, Duncan SH, Donachie GE, Holtrop G, Lo G, Farquharson FM, Wagner J, Parkhill J, Louis P, Walker AW, Flint HJ. 2020. Pivotal Roles for pH, Lactate, and Lactate-Utilizing Bacteria in the Stability of a Human Colonic Microbial Ecosystem. mSystems 5.

88. Duncan SH, Louis P, Flint HJ. 2004. Lactate-utilizing bacteria, isolated from human feces, that produce butyrate as a major fermentation product. Applied and Environmental Microbiology 70:5810–5817.

89. Wang SP, Rubio LA, Duncan SH, Donachie GE, Holtrop G, Lo G, Farquharson FM, Wagner J, Parkhill J, Louis P, Walker AW, Flint HJ. 2020. Pivotal Roles for pH, Lactate, and Lactate-Utilizing Bacteria in the Stability of a Human Colonic Microbial Ecosystem. mSystems 5.

90. Pham VT, Lacroix C, Braegger CP, Chassard C. 2016. Early colonization of functional groups of microbes in the infant gut. Environmental microbiology 18:2246–2258.

91. Yee WH, Soraisham AS, Shah VS, Aziz K, Yoon W, Lee SK, Shah PS, Andrews W, Lefebvre F, Singhal N, Bullied B, Canning R, Cronin G, Dow K, Dunn M, Harrison A, James A, Kalapesi Z, Kovacs L, Lee D, McMillan DD, Shah P, Ojah C, Piedboeuf B, Riley P, Faucher D, Rouvinez-Bouali N, Sankaran K, Seshia M, Shivananda S, Sorokan T, Synnes A. 2012. Incidence and timing of presentation of necrotizing enterocolitis in preterm infants. Pediatrics 129.

Chapter 8

The Integrated Probiotic Database: a genomic compendium of bifidobacterial health-promoting strains

Chiara Tarracchini, Martina Viglioli, Gabriele Andrea Lugli, Leonardo Mancabelli, Federico Fontana, Giulia Alessandri, Francesca Turroni, Marco Ventura, Christian Milani

The results of this chapter were published in Microbiology Spectrum, 2022 Feb; doi: 10.20517/mrr.2021.13.

Abstract

The World Health Organization defines probiotics as “live microorganisms, which when administered in adequate amounts confer a health benefit on the host.” In this framework, probiotic strains should be regarded as safe for human and animal consumption, i.e., they should possess the GRAS (Generally Recognized As Safe) status, notified by the local authorities. Consistently, strains of selected *Bifidobacterium* species are extensively used as probiotic agents to prevent and **ameliorate** a broad spectrum of human and/or animal gastrointestinal disorders.

Despite probiotic properties are often genus- or species-associated, strain-level differences in the genetic features conferring individual probiotic properties to commercialized bifidobacterial strains have not been investigated in detail.

In this study, we built a genomic database named Integrated Probiotic DataBase (IPDB), whose first iteration consists of common bifidobacterial strains used in probiotic products for which public genome sequences were available, such as members of *B. longum* subsp. *longum*, *B. longum* subsp. *infantis*, *B. bifidum*, *B. breve,* and *B. animalis* subsp. *lactis* taxa.

Furthermore, the IPDB was exploited to perform comparative genome analyses focused on genetic factors conferring structural, functional, and chemical features predicted to be involved in microbe-host and microbe-microbe interactions. Accordingly, our analyses revealed strain-level genetic differences, underlining the importance of inspecting strain-specific and outcome-specific efficacy of probiotics**.** In this context, IPDB represents a valuable resource for obtaining genetic information of well-established bifidobacterial probiotic strains.

For Supplementary Materials see the article published in Microbiome Research Reports

Introduction

The widely accepted definition of probiotics as "live microorganisms that when administered in adequate amounts confer a beneficial health effect on the host" was given by the Food and Agriculture Organization of the United Nations and the World Health Organization (FAO/WHO) in 2001 (1). Such health beneficial effects may include participation in complex carbohydrates digestion, vitamins, amino acids, and short-chain fatty acids production, antagonistic activity against intestinal bacterial pathogens, and immune system modulation (2). Nevertheless, to be considered a valuable probiotic, microbial strains must also meet specific criteria, including surviving passage through the upper digestive tract due to low pH and bile salts resistance, the ability to adhere to the human gut mucosa and to colonize the human intestine. In addition, a probiotic strain must be safe for human consumption (3).

Most microorganisms recognized to date as probiotics are Gram-positive bacteria, including members of *Enterococcus*, *Streptococcus*, *Lactobacillus*, and *Bifidobacterium* genera (4). In particular, members of this latter genus are among the main microorganisms used as probiotics in the global market (5,6). Indeed, several members of the *Bifidobacterium* genus recognized as GRAS (Generally Recognized As Safe) are widely and extensively included as live components in commercial probiotic products, either alone or in multi-strain formulations (7,8). In this context, despite bifidobacterial probiotic strains and related commercial products being accompanied by specific health-promoting claims, comparative analyses focusing on the genetic factors related to probiotic features are still lacking.

In this study, we built a genome database of the bifidobacterial strains employed in approved commercial probiotic dietary products named the Integrated Probiotic DataBase (IPDB). In detail, a total of 34 genomes corresponding to *B. longum* subsp. *longum*, *B. longum* subsp. *infantis*, *B. bifidum*, *B. breve*, and *B. animalis* subsp. *lactis* commercial probiotics were retrieved from public repositories based on extensive

literature screening and processed through an optimized bioinformatics pipeline for genes prediction and functional annotation. Further, we carried out a comparative genome analysis to identify the main shared and unique genetic features related to colonization, survival, and persistence in the gastrointestinal tract. Besides, the presence of intrinsic antimicrobial resistance (AMR) was also assessed since they could be valuable to prevent/reduce gut microbiota disorders during antibiotic treatments.

RESULTS AND DISCUSSION

The Integrated Probiotic Database (IPDB)

Bifidobacterial strain names from labels of commercially available probiotic products were identified based on comprehensive scientific literature research, and all associated publicly available genomes (complete and draft) were retrieved from NCBI (Table 1). As a result, we collected a total of 34 bifidobacterial probiotics, including four *B. longum* subsp. *infantis*, 10 *B. longum* subsp. *longum*, four *B. bifidum*, three *B. breve*, and 13 *B. animalis* subsp. *lactis* chromosomes sequences constituting the Integrated Probiotic Database (IPDB) in its first iteration. Notably, to ensure consistency in the gene prediction, all bifidobacterial genomes used in this study were re-annotated using the MEGAnnotator pipeline as described in the Methods section (9). Subsequently, the 34 bifidobacterial commercial probiotic genomes were employed to perform a comparative genome analysis to identify peculiar genetic traits possibly involved in intestinal colonization and host-microbe interaction.

All the re-annotated genome sequences, along with strain-specific functional details and information concerning the comparative analysis results, are included in the newly developed IPDB available at http://probiogenomics.unipr.it/cmu/ (direct download at http://probiogenomics.unipr.it/files/Probiotic_Bifidobacteria_DataBase.zip). Noticeably, IPDB will be expanded to include the genomes of non-bifidobacterial commercialized probiotic strains in the near future.

General genome features of the bifidobacterial strains encompassed in the IPDB

According to the genome prediction and annotation processes, we identified a number of predicted Open Reading Frames (ORFs) ranging from 2567 for *B. longum* subsp. *infantis* EVC001 to 1565 for *B. animalis* subsp. *lactis* BLC1 (Table 1).

Table 1. Publicly available bifidobacterial commercial probiotic strains included in the IPDB.

	Strain name	Assembly no.	Genome status	Genome size (Mb)	GC content (%)	No. of CDS
B. animalis* spp. *lactis	BB-12	GCA_000025245.2	Complete	1.94	60.5	1570
	BLC1	GCA_000224965.2	Complete	1.94	60.5	1565
	B420	GCA_000277325.1	Complete	1.94	60.5	1568
	BS 01	GCA_018408975.1	Draft	1.93	60.5	1728
	HN019	GCA_003606305.1	Complete	1.94	60.5	1567
	BS 05	GCA_018408985.1	Draft	2.09	60.6	1720
	MB 2409	GCA_018409015.1	Draft	1.97	60.4	1685
	Bl-04	GCA_000022705.1	Complete	1.94	60.5	1568
	Bi-07	GCA_000277345.1	Complete	1.94	60.5	1566
	ADO 11	GCA_000021425.1	Complete	1.93	60.5	1582
	BL-G101	GCA_017963615.1	Draft	1.92	60.5	1568
	BL3	GCA_002220485.1	Complete	1.94	60.5	1574
	BPL1 (CECT 8145)	GCA_000612705.1	Draft	1.96	60.4	1633
B. longum* spp. *infantis	Bi-26	GCA_004919065.2	Complete	2.61	59.3	2237
	UBBI-01	GCA_004803425.1	Draft	2.73	59.4	2462
	35624	GCA_001719085.1	Complete	2.26	60	1827
	EVC001	GCA_902167885.1	Complete	2.83	59.9	2567
B. longum* spp. *longum	BORI	GCA_003342655.1	Complete	2.31	59.9	1831
	W11	GCA_001940535.1	Draft	2.33	59.9	1886
	BL 03	GCA_018409185.1	Draft	2.35	60	2010
	DLBL 07	GCA_018409165.1	Draft	2.37	59.9	1992
	DLBL 09	GCA_018408965.1	Draft	2.37	59.8	1989
	CECT 7347 (ES1)	GCA_001050555.1	Draft	2.33	60	2019
	BB536	BAA-999 (ATCC site)	Complete	2.42	59.9	2023
	JDM301	GCA_000092325.1	Complete	2.48	59.8	2024
	KACC 91563	GCA_000219455.1	Complete	2.40	59.8	1952
	CECT 7894	GCA_016634435.1	Draft	2.29	59.9	1873
B. bifidum	PRL2010	GCA_000165905.1	Complete	2.21	62.7	1830
	BGN4	GCA_000265095.1	Complete	2.22	62.6	1792
	BF3	GCA_001281345.1	Complete	2.21	62.6	1782
	ATCC 29521	GCA_000466525.1	Draft	2.2	62.7	1846
B. breve	BR03	GCA_004319685.1	Draft	2.27	58.6	1871
	BB02	GCA_002914865.1	Draft	2.32	58.8	1983
	UBBR-01	GCA_004802595.1	Draft	2.33	58.7	2043

As previously reported, *B. longum* subsp. *infantis* showed the largest genomes among the probiotic collection, ranging between 2.83 Mb and 2.61 Mb (29), while *B.*

animalis subsp. *lactis* resulted in the taxon with the smallest genome sizes (average of 1.95 Mb).

Notably, the Average Nucleotide Identity (ANI) investigation highlighted a higher degree of genome identity among the 13 strains belonging to the *B. animalis* subsp. *lactis* species used as probiotics (average of 99.8 %), compared to all the other considered (sub)species (average of 98.1 %) (Supplementary Table S1). Although the high degrees of synteny and sequence homology between members of this taxon is well-known (ANI ~ 99.7 %) (30), 53 % of the *B. animalis* subsp. *lactis* strains showed ANI ≥99.99 %, indicating that, presumably, identical strains have been effectively deposited and commercialized with different strain names. Moreover, according to the ANI analysis, the strain *B. longum* subsp. *longum* 35624, previously misclassified as a member of the *B. longum* subsp. *infantis* taxon is still promoted commercially with an incorrect classification (Supplementary Table S1).

Overview of the commercial probiotics pan-genome. The genome sequences of the 34 bifidobacterial probiotic strains were used to predict five (sub)species-specific pan-genome profiles by classifying each strain-specific proteome into protein families named Clusters of Orthologous Groups (COGs) (Figure 1a).

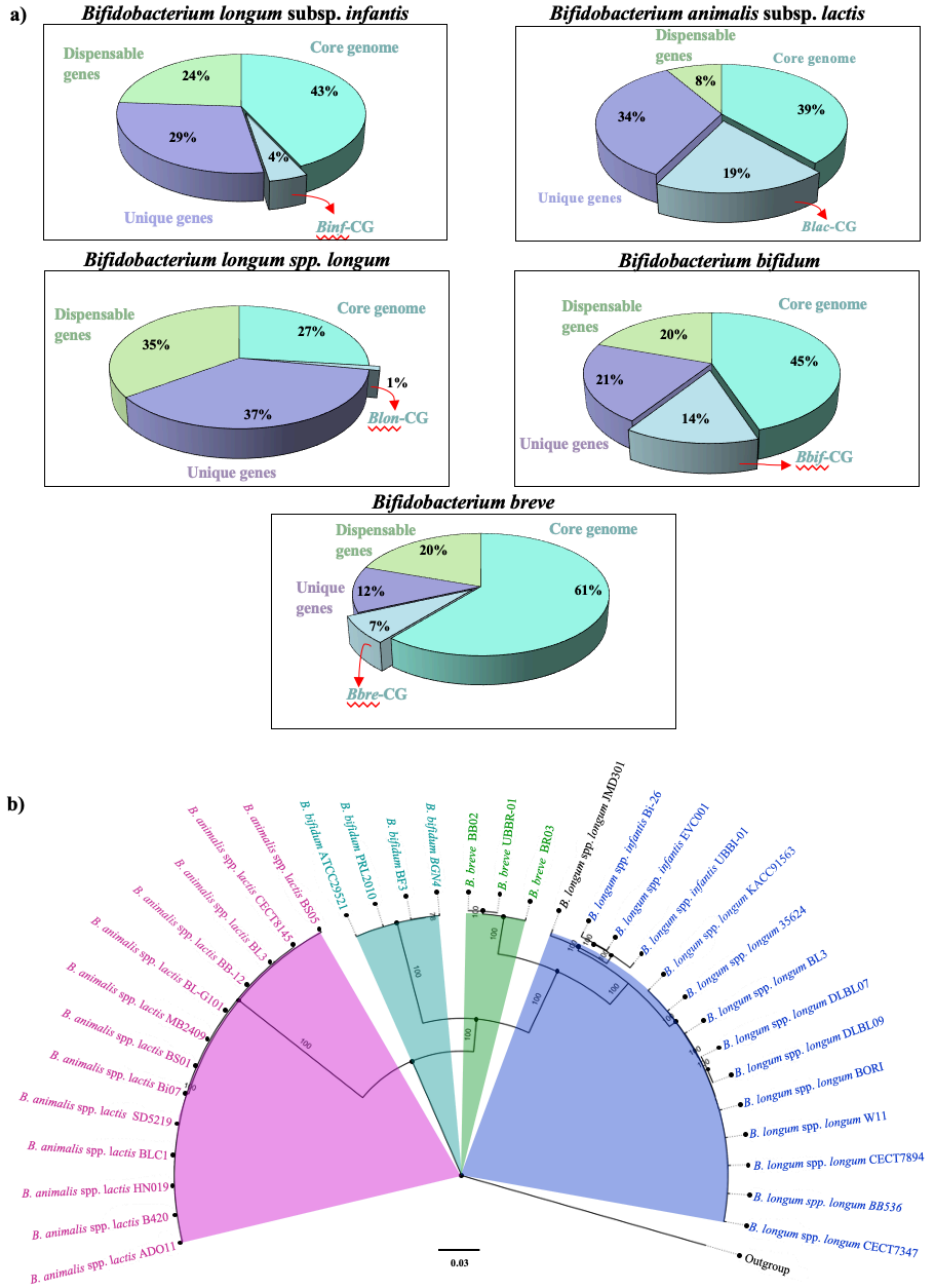


Figure 1. Pangenome of the five bifidobacterial (sub)species and phylogenetic relationships reconstruction. Panel **(a)** shows the five (sub)species-specific pangenomes profiles. The core gene pools characterizing each bifidobacterial (sub)species, i.e., *Binf*-CG, *Blon*-CG, *Bbif*-CG, *Bbre*-CG, and *Blac*-CG, were highlighted as part of each core genome. Panel **(b)** shows the phylogenomic tree based on the BPBs-CG describing the phylogenetic relationships among the 34 collected bifidobacterial probiotics. Each (sub)species-based cluster is highlighted with a different color.

Combining the obtained five (sub)species-specific pangenomes, we identified the core-genome of the bifidobacterial probiotics (BPBs-CG) by taking into account a total of 657 COGs shared by all collected bifidobacterial (sub)species (Supplementary Table S2). Similarly, five (sub)species-specific core genomes were obtained considering the COGs shared by all the strains belonging to a given sub(species) while being absent in the others. Accordingly, these latter were characterized by 150 *B. longum* subsp. *infantis* (*Binf*-CG), 90 *B. longum* subsp. *longum* (*Blon*-CG), 343 *B. bifidum (Bbif*-CG*)*, 169 *B. breve* (*Bbre*-CG), and 445 *B. animalis* subsp. *lactis* (*Blac*-CG) (sub)Species-Specific Core-genes (SSCore genes) (Figure 1a).

Notably, *B. longum* subsp. *longum* showed the lowest number of SSCore genes (ANOVA p-value < 0.05), suggesting that the evolutionary dynamics of this taxon have not led to achieving substantially unique genetic traits, while, in contrast, the phylogenetically correlated subspecies *B. longum* subsp. *infantis* showed a marked SSCore comparable to the *B. breve* species (Figure 1a, Figure 1b). Conversely, the relatively high number of SSCore genes of *Blac*-CG and *Bbif*-CG could reflect the evolutionary distance between *B. animalis* subsp. *lactis* and *B. bifidum* respect the other taxa included in this study, as pointed out by the phylogenetic reconstruction based on BPBs-CG (Figure 1a, Figure 1b).

The pan-genome analysis also revealed the strain-specific genes repertoire, i.e., Truly Unique Genes (TUGs), highlighting a variable number of TUGs ranging from 403 to 12 (average of 122.6 TUGs per genome) (Figure 1a, Supplementary Table S2). Notably, based on eggNOG analysis, bifidobacterial TUG arsenals included an average of 38.3 % of genes with general or unknown function (R/S), an average of 16.5 %, 13 %, and 9 % of genes predicted to be involved in DNA replication (M), carbohydrate (G) and amino acid (E) metabolisms, respectively, while the remaining 23.2 % were related to cell wall/membrane biogenesis (M), defense mechanism (V),

translation (J), transcription (K), and inorganic ion transport (P) (Supplementary Table S2).

Interestingly, *B. longum* subsp. *infantis* showed the highest number of TUGs (average of 322) (Supplementary Table S2). This observation indicated peculiar features that may characterize the *B. longum* subsp. *infantis* strains employed as commercial probiotics (29,31). Indeed, the relatively high degree of *B. longum* subsp. *infantis* genotype variation could be associated with the high rate of horizontal gene transfer events previously observed within this taxon (29). In contrast, *B. animalis* subsp. *lactis* exhibited the lowest number of TUGs (average of 62.7) (Supplementary Table S2), corroborating the limited genetic variability among members of this taxon (30), as evidenced by the abovementioned ANI analysis.

Distribution of host-derived glycans metabolizing capabilities providing probiotic properties

Probiotic strains can metabolize and complex dietary carbohydrates that cannot be processed by host enzymes through the production of specific glycosyl hydrolases (GHs), enhancing digestion and conferring health benefits to the host by releasing health-promoting compounds (such as SCFAs) (32). With the aim to investigate the differences in carbohydrate metabolizing capabilities of bifidobacterial probiotics, we explored the metabolic enzyme arsenal for complex carbohydrates, i.e., the glycobiome, catalyzing the breakdown of both dietary and host-derived carbohydrates. For each bifidobacterial probiotic strain, the complete glycobiome profile, including glycosyl hydrolases (GHs), glycosyl transferases (GTs), and polysaccharide lyases (PLs), was reported in the Supplementary Table S3.

Based on the CAZy database (20), we identified about 120 GHs per genome, corresponding to an average of 40.2 different GH families. In particular, 22 of the latter, including enzymes deputed to plant-derived carbohydrates metabolism as well

as GH families active on glycosidic linkages of lactose resulted to be included in the BPBs-CG, thus shared by all bifidobacterial probiotics (Supplementary Figure S1, Supplementary Table S3).

Remarkably, additional enzymes degrading host-derived glycan structures (HMOs and intestinal mucin) such as GH101 (endo-α-N-acetylgalactosaminidase), GH20 (β-hexosaminidase), GH33 (sialidase), GH129 (α-N-acetylgalactosaminidase) (www.cazy.org/) were detected in all bifidobacterial probiotics, except *B. animalis* subsp. *lactis,* and 27 % of the *B. longum* subsp. *longum* strains. Consequently, these data highlighted strain-dependent abilities of *B. longum* subsp. *longum* to digest HMOs-derived structures, and thus to promote the absorption of nutrients during infant breastfeeding (Supplementary Figure S1, Supplementary Table S3). Furthermore, the GH29 family (α-L-fucosidases) was observed highly represented in *B. bifidum* and *B. longum* subsp. *infantis* chromosomes (Supplementary Figure S1, Supplementary Table S3), while the GH84 (exo-/endo-β-N-acetylglucosaminidases) and GH89 (extracellular soluble α-N-acetylglucosaminidases) were found exclusively within *Bbif*-CG, reflecting expanded metabolic capabilities toward host-derived glycan utilization of the abovementioned taxa, compared to the other *Bifidobacterium* probiotic (sub)species (33–35).

Interestingly, members of the recently discovered GH136 family, which exert the role of extracellular lacto-N-biosidase (36), beyond being shared by all *B. bifidum* probiotics, were found in 63 % of those belonging to *B. longum* subsp. *longum* (Supplementary Figure S1, Supplementary Table S3). This observation might reveal a crucial survival strategy adopted by specific *B. longum* subsp. *longum* strains to increase their competitiveness in the infant gut ecosystem, although this subspecies is also adapted to utilize plant-derived oligosaccharides present in the adult diet (37). Overall*,* the genomes of *B. longum* subsp. *longum* showed the highest number of accessory GH genes (Figure 2, Supplementary Table S3). Indeed, within the

chromosomes of this taxon, a total of 24 GH families were found from 90 % to 9 % of the strains, in comparison of only 2-8 GH families constituting the accessory GH arsenal of the other considered probiotic (sub)species (Figure 2, Supplementary Table S3). In particular, GH families involved in the degradation of HMOs and host glycan structures, i.e., GH129, GH136, GH85 (endo-β-N-acetylglucosaminidase), and GH29, were found in 72.7 %, 63.6 %, 45.5 %, and 9.1 % of the probiotic strains belonging to *B. longum* subsp. *longum*.

Although carbohydrate utilization capabilities are often associated with (sub)species-specific features, IPDB analyses reported differences in carbohydrate-metabolizing enzymes between commercialized probiotics of the same (sub)species. Such differences can have functional and ecological implications worthy of being taken into account for probiotic formulation and consumption.

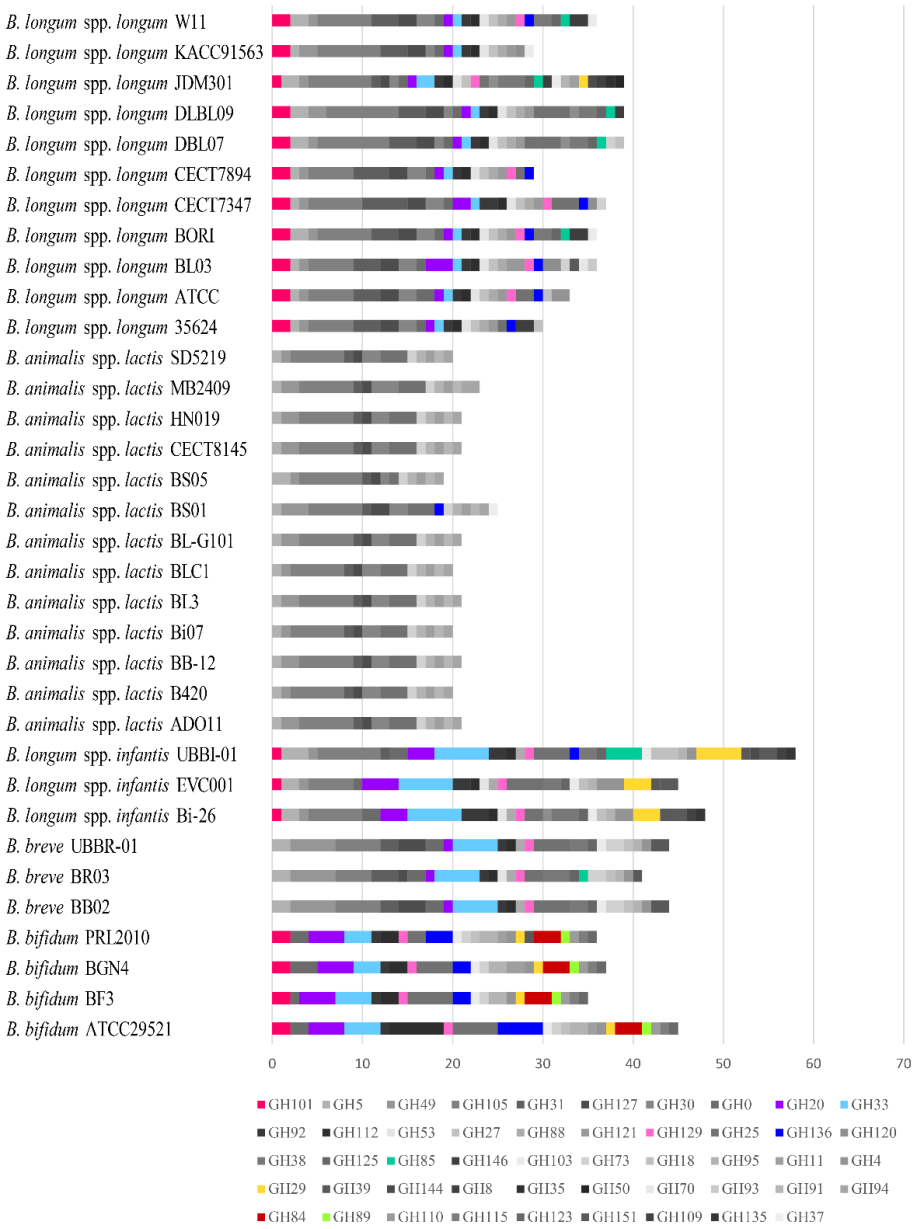


Figure 2. Accessory GH profiles of the bifidobacterial probiotics. For each bifidobacterial probiotic strain, the occurrence of the accessory GHs, i.e., GHs shared by a subset of the considered probiotic strains, are depicted through a bar-plot graph. The accessory GH families active on the host-derived glycans mentioned in the text are highlighted with different colors.

Extracellular structures involved in microbe-host interactions

Bacterial extracellular appendages, such as pili or fimbriae, are long and non-flagellar structures strategically localized to the cell surface to promote bacterial adhesion in the gut, simultaneously impacting microbe-host dialogue (38,39). In the *Bifidobacterium* genus, sortase-dependent (SD) pili (types I and II), collectively representing the SD fimbriome, as well as type IV pili, have been previously described (26). While these latter are highly conserved among bifidobacterial genomes, the SD pili showed a considerable variability (40). According to these notions, we explored the SD pili-encoding genes arsenal of the 34 collected bifidobacterial probiotic strains exploiting a custom database built in the contest of a previous study (26).

Overall, SD pilus gene clusters, composed of a sortase-encoding gene for assembling pilus subunits and two pilin subunit-encoding genes, were found in 91 % of the inspecting genomes (Figure 3, Supplementary Table S4). Interestingly, while genome sequences of *B. longum* susp. *infantis* appear unable to encode this type of pili, probiotic strains belonging to *B. bifidum* possessed the highest number of SD pili-encoding clusters (Figure 3, Supplementary Table S4).

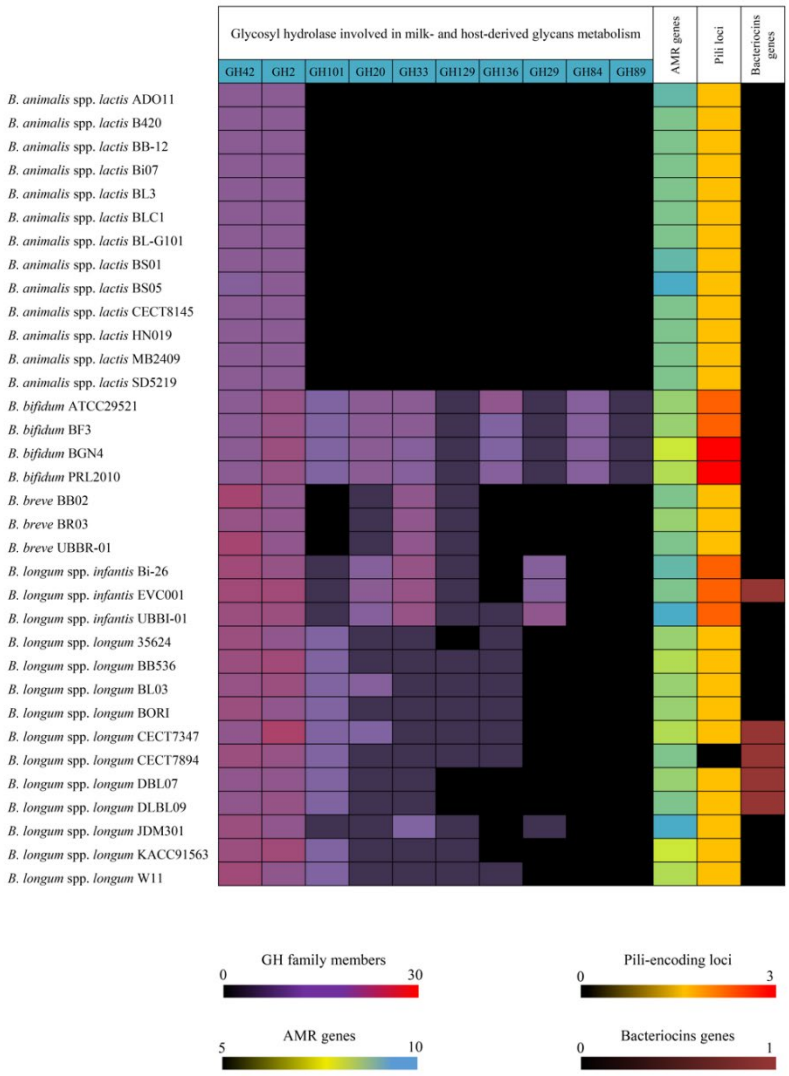


Figure 3. Occurrence of genetic probiotic features in the bifidobacterial strains. For each of the 34 considered bifidobacterial probiotic strain, the heat map shows the predicted number of glycosyl hydrolase enzymes involved in host glycan metabolisms, antimicrobial resistance determinants, pili- and bacteriocins-encoding genes, bile salt hydrolases, and exopolysaccharides (EPSs)-encoding loci.

In particular, the genomes of *B. bifidum* strains BGN4 and PRL2010 showed three SD pili loci, thus suggesting putative improved adherence and persistence features. Furthermore, a diverse array of genes required for the production of SD pili was observed between probiotic strains belonging to the same (sub)species. In particular,

B. longum subsp. *longum* showed a variable number of SD pili, ranging from 0 to 2 (Figure 3, Supplementary Table S4), highlighting possible different abilities to colonize and persist in the human gastrointestinal tract.

Overall, these data collected in the IPDB, in addition to (sub)species-specific features, highlighted considerable strain-level variabilities in the environment interaction structures that could therefore determine different individual probiotic properties.

Production of bacteriocins by commercial bifidobacterial probiotics

In addition to external structures, bifidobacteria exploit molecule-based systems to compete for intestinal colonization directly. Although the inhibitory activity of bifidobacteria could partially derive from the production of organic acids, it is hypothesized that some members of the *Bifidobacterium* genus can produce antimicrobial molecules such as bacteriocins (41,42). These latter are ribosomally-synthesized peptides with antimicrobial activities against other bacteria, either belonging to the same species or even across genera (43,44). Consequently, these compounds are regarded as a probiotic trait contributing to higher niche competitiveness and inhibition of intestinal pathogens (45). For this reason, we investigated the occurrence of bacteriocins-encoding genes among the 34 bifidobacterial probiotics using the BAGEL4 database (27).

As a result, a total of five potential bacteriocin genes were predicted to be codified by *B. longum* subsp. *infantis* and *B. longum* subsp. *longum* probiotic strains. In particular, a Class I lantibiotic (BLD_1648) was found in *B. longum* subsp. *infantis* EVC001, and in four members of *B. longum* subsp. *longum* taxon, i.e., strains CECT7347, CECT7894, DLBL07, and DLBL09 (Figure 3, Supplementary Table S4).

Based on this *in silico* analyses results, only a limited number of *Bifidobacterium* species encode for bacteriocins, and intra-(sub)species variabilities have been found when comparing different strains. In particular, only (certain) strains belonging to *B. longum* subsp. *longum* and *B. longum* subsp. *infantis* showed strain-specific abilities in producing antimicrobial compounds, which may facilitate the introduction of the (probiotic) producer into an established niche by directly inhibiting competing strains or pathogens. Thus, these findings evidenced by analysis of the IPDB reinforce the need for a precise assessment of desirable probiotic properties, such as bacteriocins production, at a strain-specific level.

Antibiotics resistance prediction and their distribution among commercial bifidobacterial probiotics

Probiotics are specifically selected not to carry AntiMicrobial Resistance (AMR), with particular attention to AMR determinants located in the proximity of transposable elements or falling inside (integrated) bacterial plasmids, which could contribute to the spread of AMR (46,47).

Notably, AMR determinants surveys across the *Bifidobacterium* genus revealed that, except for tetracyclines resistance (*tet* genes) in specific cases, the resistance phenotypes are independent of the presence of particular genes, or they do not fall in genomic regions involved in horizontal gene transfer events. Hence, they rarely represent a risk for transfer to unrelated pathogenic or potentially pathogenic bacteria (48,49). On the other hand, AMR can enhance the survival of the probiotics in the presence of antimicrobial compounds due to medical treatments, thus constituting a beneficial feature (50).

In this contest, the collected 34 bifidobacterial probiotic strains were inspected for putative antibiotic resistance determinants. Even if our *in silico* analysis remains only predictive, such an approach can provide indications for further *in vitro* validations.

As a result, an average of 7.6 AMR genes per chromosome were identified (Supplementary Table S5). Among these, a putative ATP-binding cassette (ABC) transporter that exports macrolides (ARO:3000535), as well as putative rifampicin (ARO:3004480), fosfomycin (ARO:3003785), and mupirocin (ARO:3003730) resistances were found shared by all probiotics, while a gene conferring resistance to cationic antimicrobial peptides (ARO:3003577) shared by 76 % of bifidobacterial probiotics strains (Figure 3, Supplementary Table S5). Notably, prediction of the putative transporter specificity was assessed with manual validation employing the Transporter Classification DataBase (TCDB) (24).

In addition, also (sub)species-specific AMR genes were observed, including genes putatively involved in resistance against several different classes of antibiotics, i.e., multidrug efflux transporter, within *Blon*-CG (ARO:3000816) and *Bbre*-CG (ARO:3002813), and genes conferring putative resistance to mycinamicin (ARO:3001301) and tetracycline antibiotics (ARO:3003980, ARO:3000194) in *Blac*-CG (Figure 3, Supplementary Table S5).

Focusing on the unique gene pools, which could be horizontally acquired, strain-specific AMR genes were found in four out of the 34 screened probiotic strains (Figure 3, Supplementary Table S5). In particular, *B. animalis* subsp. *lactis* BS05, showing the lowest ANI value, i.e., 99.3 %, when compared to the other genomes of the same species, was predicted to encode a mosaic tetracycline resistance gene (*tetW/N/W*, ARO:3004442) and a gene possibly involved in resistances to carbapenems, rifamycin, and peptide antibiotics (ARO:3005059) (Figure 3, Supplementary Table S5). Moreover, two different macrolide resistance systems (ARO:3000616 and ARO:3004626) were noticed in the *B. longum* subsp. *longum* CECT7894 and JDM301 strains, respectively, whereas a generic antibiotic efflux pump (ARO:3000838) was observed in the strain DLBL09 (Figure 3, Supplementary Table S5).

Based on the data collected in the IPDB, bifidobacterial probiotics appear to possess a relatively low acquired resistance compared to members of the *Enterococcus* and *Lactobacillus* genus used as probiotics in humans and farm animals (51), for which resistance to a wide range of antimicrobials carried on plasmids or in the proximity of conjugative transposons have been identified (52–54). Nevertheless, strain-specific AMR determinants have been observed, highlighting the need for case-by-case assessments.

Screening of additional genetic features involved in colonization and persistence

To further characterize the 34 bifidobacterial probiotic strains included in the database, attributes including bile salt tolerance mediated by bile salt hydrolases (BSH), production of extracellular polysaccharides (EPSs), and the presence of putative virulence factors were investigated. Notably, the ability to hydrolyze bile salts is often regarded as a desirable feature for probiotic strain selection since it can promote probiotic fitness and colonization by detoxifying bile (55). According to *in silico* analyses, bile salt hydrolase activity has been predicted for all the 34 bifidobacterial strains (Supplementary Table S6), resulting to be a widespread trait among the bifidobacterial probiotics. Furthermore, screening of potential virulence-related genes revealed the presence of homologous genes associated to surface carbohydrates polymers and response regulator proteins which typically mediate the interaction with the surrounding environment (Supplementary Table S7). However, such structures are not recognized as harmful. Instead, they are well-known to participate in the host-microbe dialogue underlying and supporting the claimed probiotic effects (56). Consistently, the analysis revealed the absence of genes associated with clear detrimental effects, remarking the safe use of bifidobacterial strains as probiotics.

Among the interesting and attractive characteristics of probiotic strains, the production of Exopolysaccharides (EPSs) has grasped the attention because of its important role in maintaining commensalism between human host and (bifido)bacteria as well as for their putative health-promoting properties (57,58). EPSs are extracellular carbohydrate polymers, for whose biosynthesis a gene cluster including a priming glycosyltransferase (pGTF) and additional genes, such as ABC transporters, subunit polymerization enzymes, and carbohydrate precursor biosynthesis/modification enzymes, are required (59). In particular, pGTF is an essential enzyme that catalyzes the first step of the EPSs biosynthetic pathway (59). In this context, the 34 bifidobacterial probiotics were explored for EPS loci employing well-known pGTF gene sequences as molecular indicators, as previously performed (60,61).

Accordingly, the production of EPSs was predicted in all bifidobacterial chromosomes except for those belonging to *B. bifidum* species. Specifically, the presence of two highly conserved EPS loci were observed in each *B. animalis* subsp. *lactis* probiotic (Supplementary Table S6), while a single EPS-producing locus with a significant intra-(sub)species variability was detected among probiotic strains belonging to *B. longum* subsp. *longum* and *B. longum* subsp. *infantis* taxa (Supplementary Table S6). The precise location of the pGTFs predicted in each bifidobacterial genome is reported in Supplementary Table S6.

Conclusion

Because of their safety, functional, and technological characteristics, various members of the *Bifidobacterium* genus have been commercially available to and steadily used as probiotic bacteria.

In this study, we constructed the first iteration of a genomic database named Integrated Probiotic DataBase (IPDB) encompassing 34 publicly available strains of *B. bifidum*, *B. longum* subsp. *longum*, *B. longum* subsp. *infantis*, *B. breve* and *B. animalis* subsp. *lactis* (sub)species used in commercialized health-promoting supplements. The collected genome sequences were re-analyzed using an updated bioinformatics pipeline, and all the acquired genetic and functional information was included in the IPDB. Comparative genome analyses, in addition to genetic determinants shared by all the members of a species, revealed the existence of a range of strain-unique features possibly related to probiotic activities.

In particular, the greater number of host glycans-metabolizing and pili-encoding genes found in the genome sequences of *B. bifidum* and *B. longum* subsp. *infantis* (sub)species reflect their higher capability to colonize and persist in the human gastrointestinal tract as well as in those of lactating infants. On the other hand, strain-specific host-derived glycans metabolic machinery was deployed by some strains of *B. longum* subsp. *longum*, reflecting intra-(sub)species differences in enhancing digestion and absorption of nutrients in breastfed infants. Moreover, strain-dependent differences in bacteriocins production, EPSs biosynthesis, and antibiotic resistances were noticed not only among probiotic species, but potentially among strains of the same species. Accordingly, strain-specific gene arsenals deserve attention since they can be correlated with profound different ecological behavior in the intestinal environment and the dialogue with the host, thus leading to different probiotic outcomes. As a result, accurate strain-level information about probiotic products should now be considered necessary to allow consumers to obtain precise evidence behind the claimed beneficial effects of each probiotic.

In this context, the IPDB represents a novel intriguing instrument to rapidly access the genome content of common bifidobacterial probiotic strains, assisting in drawing the connection between probiotics, gut microbiome, and beneficial effects to the host.

METHODS

Genome Sequences of bifidobacterial commercial strains

In accordance with scientific literature surveys, publicly available chromosomal sequences of 34 bifidobacterial strains used in commercial dietary probiotic products were retrieved from the National Center for Biotechnology Information (NCBI) public database (Table 1). In order to ensure a consistent genomic analysis, ORFs from each bifidobacterial genome sequence were re-predicted and annotated using the most recent release of the MEGAnnotator pipeline (9). In detail, contigs greater than 1000 bp were employed to predict protein-encoding ORFs through Prodigal v2.0 (Linux command line './prodigal -f gff -a [protein_translation_to_selected_file] -i [input_filename.fasta] -o [output_filename]") (10). Predicted ORFs were then functionally annotated using RAPSearch2 (reduced alphabet-based protein similarity search) (cutoff e-value of 1×10^{-5} and minimum alignment length 20) employing the NCBI reference sequences (RefSeq) database (11) together with hidden Markov model profile (HMM) searches (http://hmmer.org/) against the manually curated Pfam-A database (cutoff e-value of 1×10^{-10}). Then, tRNA genes were detected through tRNAscan-SE v1.4 (12), while rRNA genes were identified using RNAmmer v1.2 (13).

Comparative Genomic Analysis

All 34 genome sequences of *Bifidobacterium* members were employed for a pan-genome analysis using the Pangenome Analysis Pipeline (PGAP) v1.1 (http://pgap.sf.net) (14). The predicted proteome of each bifidobacterial genome was classified into functional gene clusters through the gene family (GF) method, consisting of pairwise protein-similarity search using blast software v2.2.28+ (cutoff e-value of 1×10^{-10} and exhibiting at least 50% identity across at least 80% of both

protein sequences). The obtained data were used to assign proteins to so-called Clusters of Orthologous Groups (COGs) employing MCL (graph-theory-based Markov clustering algorithm) (15). A pan-genome profile was then built using a presence/absence matrix encompassing all COGs identified in the analysed genomes (Linux command line './PGAP.pl --strains [input_strain_list] --input input_path/ --output output_path/ --thread 20 --identity 0.5 --coverage 0.8 --cluster --method GF --evolution --pangenome'). Subsequently, the core genome of commercial bifidobacterial strains was obtained by selecting the protein families shared between all genomes, while truly unique genes (TUGs) of a given genome were identified based on those protein families that are not present in other bifidobacterial chromosomes. Functional annotation of each TUG arsenal was accomplished employing the eggNOG database (16). Each pairwise average nucleotide identity (ANI) was calculated using the program fastANI (17).

Phylogenomic analysis

In order to disentangle the phylogenetic relationships between the 34 collected bifidobacterial probiotic strains, the concatenated sequence of amino acids belonging to the core genome of each bifidobacterial strain was aligned using the MAFFT software (18). The resulting phylogenetic tree was built using the neighbor-joining method through the ClustalW v2.1 program (18), and the graphical viewer of phylogenetic trees FigTree v1.4 (http://tree.bio.ed.ac.uk/software/figtree/) was used to its visual representation.

Glycobiome prediction and identification of genes conferring antimicrobial resistance

The genome sequences of the publicly available 34 bifidobacterial probiotic strains were subjected to assessment of genes encoding for glycosyl hydrolase (GH),

glycosyl transferase (GT), and polysaccharides lyase (PL) enzymes through sequence similarity search in the carbohydrate-active enzyme (CAZy) database (20) using HMMER v3.3 (21) (cutoff e-value of 1×10^{-15}) and BLASTP analysis (22) (cutoff e-value of 1×10^{-10}).

The proteome of each bifidobacterial probiotics genome was also screened for the presence of bacterial antimicrobial resistance based on sequence similarity to genes classified in the Comprehensive Antibiotic Resistance Database (CARD) (23) (BLASTP cutoff e-value of 1×10^{-5}). Outcomes were then manually validated to eliminate possible false positives. Moreover, Transporter Classification DataBase (TCDB) (24)was employed to assess the putative transporter specificity.

Identification of sortase-dependent (SD) pilus-encoding loci, and bacteriocins-encoding genes

SD pilus-encoding loci (type I and type II pili) were identified through homology search tool RAPsearch (cutoff E value of 1×10^{-5}; minimal alignment length 20) (25) exploiting the custom sortase-dependent pilus genes database previously built (26). Then, a detailed manual inspection was performed to identify complete pilus gene clusters.

Likewise, bacteriocin-encoding genes were detected using RAPsearch analysis (cutoff E value of 1×10^{-5}; minimal alignment length 20) employing the BAGEL4 database (27).

Assessment the genetic background for exopolysaccharides, virulence, and bile salt hydrolases production

In order to identify the loci encoding exopolysaccharides (EPSs), the protein sequences of well-known priming glycosyltransferases (pGTFs) were retrieved from NCBI database and were used to inspect the 34 bifidobacterial genome sequences.

Subsequently, for each bifidobacterial chromosome, the genomic regions flanking the putative pGTF were investigated to identify EPS-encoding key genes (such as glycosyltransferases, flippases, ABC transporters, and carbohydrate precursor biosynthesis/modification enzymes). The presence of putative virulence genes and bile salt hydrolases were identified through sequence similarity (homology) search in the Virulence Factor Database (VFDB) (28) and in the protein sequence NCBI database, respectively (cutoff E value of 1×10^{-5}). Thus, the resulting hits were manually inspected to remove false positives.

Statistical Analyses. All statistical analyses were computed using SPSS software (www.ibm.com/software/it/analytics/spss/).

Acknowledgments

Part of this research is conducted using the High Performance Computing (HPC) facility of the University of Parma.

Authors' Contributions

Data analysis and manuscript writing: Tarracchini C.

Data acquisition: Viglioli M.

Data curation and data analysis: Lugli G.A., Mancabelli L., Fontana F., Alessandri G.

Conceptualization, Supervision and manuscript editing: Turroni F., Milani C., Ventura M.

Availability of Data and Materials

The IPDB can be accessed at http://probiogenomics.unipr.it/cmu/ (direct download at http://probiogenomics.unipr.it/files/Probiotic_Bifidobacteria_DataBase.zip).

Financial Support and Sponsorship

We thank GenProbio Srl for the financial support of the Laboratory of Probiogenomics.

Ethical Approval and Consent to Participate

Not applicable.

Conflicts of Interest

All authors declared that there are no conflicts of interest.

Consent for Publication

Not applicable.

Copyright

References

1. Hill C, Guarner F, Reid G, Gibson GR, Merenstein DJ, Pot B, et al. The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic. Nature Reviews Gastroenterology & Hepatology 2014 11:8. 2014 Jun 10;11(8):506–14.

2. Aditya A, Peng M, Young A, Biswas D. Antagonistic Mechanism of Metabolites Produced by Lactobacillus casei on Lysis of Enterohemorrhagic Escherichia coli. Frontiers in microbiology [Internet]. 2020 Nov 23 [cited 2021 Dec 9];11. Available from: https://pubmed.ncbi.nlm.nih.gov/33329433/

3. Tomasik PJ, Tomasik P. Probiotics and Prebiotics. Cereal Chemistry [Internet]. 2003 Mar 1 [cited 2021 Oct 19];80(2):113–7. Available from: https://onlinelibrary.wiley.com/doi/full/10.1094/CCHEM.2003.80.2.113

4. Koutsoumanis K, Allende A, Alvarez-Ordóñez A, Bolton D, Bover-Cid S, Chemaly M, et al. Update of the list of QPS-recommended biological agents intentionally added to food or feed as notified to EFSA 14: suitability of taxonomic units notified to EFSA until March 2021. EFSA journal European Food Safety Authority [Internet]. 2021 Jul 1 [cited 2021 Dec 13];19(7). Available from: https://pubmed.ncbi.nlm.nih.gov/34257732/

5. Scheinbach S. Probiotics: functionality and commercial status. Biotechnology advances [Internet]. 1998 May [cited 2021 Dec 9];16(3):581–608. Available from: https://pubmed.ncbi.nlm.nih.gov/14538145/

6. Papizadeh M, Rohani M, Nahrevanian H, Javadi A, Pourshafie MR. Probiotic characters of Bifidobacterium and Lactobacillus are a result of the ongoing gene acquisition and genome minimization evolutionary trends. Microbial pathogenesis [Internet]. 2017 Oct 1 [cited 2021 Dec 9];111:118–31. Available from: https://pubmed.ncbi.nlm.nih.gov/28826768/

7. McFarland L v. Efficacy of Single-Strain Probiotics Versus Multi-Strain Mixtures: Systematic Review of Strain and Disease Specificity. Digestive diseases and sciences [Internet]. 2021 Mar 1 [cited 2021 Dec 9];66(3):694–704. Available from: https://pubmed.ncbi.nlm.nih.gov/32274669/

8. Ouwehand AC, Invernici MM, Furlaneto FAC, Messora MR. Effectiveness of Multistrain Versus Single-strain Probiotics: Current Status and Recommendations for the Future. Journal of clinical gastroenterology [Internet]. 2018 Nov 1 [cited 2021 Dec 9];52:S35–40. Available from: https://pubmed.ncbi.nlm.nih.gov/29734210/

9. Lugli GA, Milani C, Mancabelli L, van Sinderen D, Ventura M. MEGAnnotator: a user-friendly pipeline for microbial genomes assembly and annotation. FEMS microbiology letters [Internet]. 2016 Apr 1 [cited 2021 Dec 9];363(7). Available from: https://pubmed.ncbi.nlm.nih.gov/26936607/

10. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics [Internet]. 2010 Mar 8 [cited 2021 Dec 9];11. Available from: https://pubmed.ncbi.nlm.nih.gov/20211023/

11. Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics (Oxford, England) [Internet]. 2012 Jan [cited 2021 Dec 9];28(1):125–6. Available from: https://pubmed.ncbi.nlm.nih.gov/22039206/

12. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic acids research [Internet]. 1997 Mar 1 [cited 2021 Dec 9];25(5):955–64. Available from: https://pubmed.ncbi.nlm.nih.gov/9023104/

13. Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic acids research [Internet]. 2007 May [cited 2021 Dec 9];35(9):3100–8. Available from: https://pubmed.ncbi.nlm.nih.gov/17452365/

14. Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J. PGAP: pan-genomes analysis pipeline. Bioinformatics (Oxford, England) [Internet]. 2012 Feb [cited 2021 Dec 9];28(3):416–8. Available from: https://pubmed.ncbi.nlm.nih.gov/22130594/

15. Enright AJ, van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic acids research [Internet]. 2002 Apr 1 [cited 2021 Dec 9];30(7):1575–84. Available from: https://pubmed.ncbi.nlm.nih.gov/11917018/

16. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic acids research [Internet]. 2016 [cited 2021 Dec 29];44(D1):D286–93. Available from: https://pubmed.ncbi.nlm.nih.gov/26582926/

17. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature communications [Internet]. 2018 Dec 1 [cited 2021 Dec 9];9(1). Available from: https://pubmed.ncbi.nlm.nih.gov/30504855/

18. Katoh K, Misawa K, Kuma KI, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research [Internet]. 2002 Jul 15 [cited 2021 Dec 9];30(14):3059–66. Available from: https://pubmed.ncbi.nlm.nih.gov/12136088/

19. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, et al. Multiple sequence alignment with the Clustal series of programs. Nucleic acids research [Internet]. 2003 Jul 1 [cited 2021 Dec 9];31(13):3497–500. Available from: https://pubmed.ncbi.nlm.nih.gov/12824352/

20. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic acids research [Internet]. 2014 Jan 1 [cited 2021 Dec 9];42(Database issue). Available from: https://pubmed.ncbi.nlm.nih.gov/24270786/

21. Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. Bioinformatics (Oxford, England) [Internet]. 2013 Oct 1 [cited 2021 Dec 9];29(19):2487–9. Available from: https://pubmed.ncbi.nlm.nih.gov/23842809/

22. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology [Internet]. 1990 [cited 2021 Dec 9];215(3):403–10. Available from: https://pubmed.ncbi.nlm.nih.gov/2231712/

23. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic acids research [Internet]. 2020 Jan 1 [cited 2021 Dec 9];48(D1):D517–25. Available from: https://pubmed.ncbi.nlm.nih.gov/31665441/

24. Saier MH, Reddy VS, Tsu B v., Ahmed MS, Li C, Moreno-Hagelsieb G. The Transporter Classification Database (TCDB): recent advances. Nucleic acids research [Internet]. 2016 [cited 2021 Dec 15];44(D1):D372–9. Available from: https://pubmed.ncbi.nlm.nih.gov/26546518/

25. Ye Y, Choi JH, Tang H. RAPSearch: a fast protein similarity search tool for short reads. BMC bioinformatics [Internet]. 2011 May 15 [cited 2021 Dec 9];12. Available from: https://pubmed.ncbi.nlm.nih.gov/21575167/

26. 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 [Internet]. 2017 [cited 2021 Dec 9];83(19). Available from: https://pubmed.ncbi.nlm.nih.gov/28754709/

27. van Heel AJ, de Jong A, Song C, Viel JH, Kok J, Kuipers OP. BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. Nucleic acids research [Internet]. 2018 Jul 2 [cited 2021 Dec 9];46(W1):W278–81. Available from: https://pubmed.ncbi.nlm.nih.gov/29788290/

28. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, et al. VFDB: a reference database for bacterial virulence factors. Nucleic acids research [Internet]. 2005 Jan 1 [cited 2022 Jan 11];33(Database issue). Available from: https://pubmed.ncbi.nlm.nih.gov/15608208/

29. Tarracchini C, Milani C, Lugli GA, Mancabelli L, Fontana F, Alessandri G, et al. Phylogenomic disentangling of the Bifidobacterium longum subsp. infantis taxon. Microbial genomics [Internet]. 2021 Jul 28 [cited 2021 Dec 9];7(7). Available from: https://pubmed.ncbi.nlm.nih.gov/34319225/

30. Blanco-Míguez A, Gutiérrez-Jácome A, Fdez-Riverola F, Lourenço A, Sánchez B. A peptidome-based phylogeny pipeline reveals differential peptides at the strain level within Bifidobacterium animalis subsp. lactis. Food microbiology [Internet]. 2016 Dec 1 [cited 2021 Dec 9];60:137–41. Available from: https://pubmed.ncbi.nlm.nih.gov/27554155/

31. Underwood MA, German JB, Lebrilla CB, Mills DA. Bifidobacterium longum subspecies infantis: champion colonizer of the infant gut. Pediatric research [Internet]. 2015 Jan 10 [cited 2021 Dec 9];77(1–2):229–35. Available from: https://pubmed.ncbi.nlm.nih.gov/25303277/

32. Tan J, McKenzie C, Potamitis M, Thorburn AN, Mackay CR, Macia L. The role of short-chain fatty acids in health and disease. Advances in immunology [Internet]. 2014 [cited 2021 Dec 9];121:91–119. Available from: https://pubmed.ncbi.nlm.nih.gov/24388214/

33. Abdelhamid AG, El-Dougdoug NK. Comparative genomics of the gut commensal Bifidobacterium bifidum reveals adaptation to carbohydrate utilization. Biochemical and biophysical research communications [Internet]. 2021 Apr 2 [cited 2021 Dec 9];547:155–61. Available from: https://pubmed.ncbi.nlm.nih.gov/33610915/

34. Sela DA, Garrido D, Lerno L, Wu S, Tan K, Eom HJ, et al. Bifidobacterium longum subsp. infantis ATCC 15697 α-fucosidases are active on fucosylated human milk oligosaccharides. Applied and environmental microbiology [Internet]. 2012 Feb [cited 2021 Dec 9];78(3):795–803. Available from: https://pubmed.ncbi.nlm.nih.gov/22138995/

35. Katoh T, Ojima MN, Sakanaka M, Ashida H, Gotoh A, Katayama T. Enzymatic Adaptation of Bifidobacterium bifidum to Host Glycans, Viewed from Glycoside Hydrolyases and Carbohydrate-Binding Modules. Microorganisms [Internet]. 2020 Apr 1 [cited 2021 Dec 9];8(4). Available from: https://pubmed.ncbi.nlm.nih.gov/32231096/

36. Sakurama H, Kiyohara M, Wada J, Honda Y, Yamaguchi M, Fukiya S, et al. Lacto-N-biosidase encoded by a novel gene of Bifidobacterium longum subspecies longum shows unique substrate specificity and requires a designated chaperone for its active expression. The Journal of biological chemistry [Internet]. 2013 Aug 30 [cited 2021 Dec 9];288(35):25194–206. Available from: https://pubmed.ncbi.nlm.nih.gov/23843461/

37. 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 [Internet]. 2018 Dec 1 [cited 2021 Dec 9];8(1). Available from: https://pubmed.ncbi.nlm.nih.gov/29311585/

38. Turroni F, Serafini F, Foroni E, Duranti S, Motherway MOC, 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 of the United States of America [Internet]. 2013 Jul 2 [cited 2021 Dec 9];110(27):11151–6. Available from: https://pubmed.ncbi.nlm.nih.gov/23776216/

39. Nishiyama K, Yokoi T, Sugiyama M, Osawa R, Mukai T, Okada N. Roles of the Cell Surface Architecture of Bacteroides and Bifidobacterium in the Gut Colonization. Frontiers in microbiology [Internet]. 2021 Oct 14 [cited 2021 Dec 9];12. Available from: https://pubmed.ncbi.nlm.nih.gov/34721360/

40. Alessandri G, van Sinderen D, Ventura M. The genus bifidobacterium: From genomics to functionality of an important component of the mammalian gut microbiota running title: Bifidobacterial adaptation to and interaction with the host. Computational and Structural Biotechnology Journal [Internet]. 2021 Jan 1 [cited 2021 Dec 10];19:1472. Available from: /pmc/articles/PMC7979991/

41. Lievin V, Peiffer I, Hudault S, Rochat F, Brassart D, Neeser JR, et al. Bifidobacterium strains from resident infant human gastrointestinal microflora exert antimicrobial activity. Gut [Internet]. 2000 [cited 2021 Dec 13];47(5):646–52. Available from: https://pubmed.ncbi.nlm.nih.gov/11034580/

42. de Niederhäusern S, Camellini S, Sabia C, Iseppi R, Bondi M, Messi P. Antilisterial Activity of Bacteriocins Produced by Lactic Bacteria Isolated from Dairy Products. Foods (Basel, Switzerland) [Internet]. 2020 Dec 1 [cited 2021 Dec 13];9(12). Available from: https://pubmed.ncbi.nlm.nih.gov/33260958/

43. Kanmani P, Satish Kumar R, Yuvaraj N, Paari KA, Pattukumar V, Arul V. Probiotics and its functionally valuable products-a review. Critical reviews in food science and nutrition [Internet]. 2013 Jan [cited 2021 Dec 9];53(6):641–58. Available from: https://pubmed.ncbi.nlm.nih.gov/23627505/

44. Lievin V, Peiffer I, Hudault S, Rochat F, Brassart D, Neeser JR, et al. Bifidobacterium strains from resident infant human gastrointestinal microflora exert antimicrobial activity. Gut [Internet]. 2000 [cited 2021 Dec 9];47(5):646–52. Available from: https://pubmed.ncbi.nlm.nih.gov/11034580/

45. O'Shea EF, Cotter PD, Stanton C, Ross RP, Hill C. Production of bioactive substances by intestinal bacteria as a basis for explaining probiotic mechanisms: bacteriocins and conjugated linoleic acid. International journal of food microbiology [Internet]. 2012 Jan 16 [cited 2021 Dec 9];152(3):189–205. Available from: https://pubmed.ncbi.nlm.nih.gov/21742394/

46. Gama JA, Zilhão R, Dionisio F. Impact of plasmid interactions with the chromosome and other plasmids on the spread of antibiotic resistance. Plasmid [Internet]. 2018 Sep 1 [cited 2021 Dec 9];99:82–8. Available from: https://pubmed.ncbi.nlm.nih.gov/30240700/

47. Lupski JR. Molecular mechanisms for transposition of drug-resistance genes and other movable genetic elements. Reviews of infectious diseases [Internet]. 1987 [cited 2021 Dec 9];9(2):357–68. Available from: https://pubmed.ncbi.nlm.nih.gov/3035697/

48. Kiwaki M, Sato T. Antimicrobial susceptibility of Bifidobacterium breve strains and genetic analysis of streptomycin resistance of probiotic B. breve strain Yakult. International journal of food microbiology [Internet]. 2009 Sep 15 [cited 2021 Dec 9];134(3):211–5. Available from: https://pubmed.ncbi.nlm.nih.gov/19616336/

49. Sato T, Iino T. Genetic analyses of the antibiotic resistance of Bifidobacterium bifidum strain Yakult YIT 4007. International journal of food microbiology [Internet]. 2010 Feb 28 [cited 2021 Dec 9];137(2–3):254–8. Available from: https://pubmed.ncbi.nlm.nih.gov/20051305/

50. Gueimonde M, Sánchez B, de los Reyes-Gavilán CG, Margolles A. Antibiotic resistance in probiotic bacteria. Frontiers in microbiology [Internet]. 2013 [cited 2021 Dec 9];4(JUL). Available from: https://pubmed.ncbi.nlm.nih.gov/23882264/

51. Franz CMAP, Huch M, Abriouel H, Holzapfel W, Gálvez A. Enterococci as probiotics and their implications in food safety. International journal of food microbiology [Internet]. 2011 Dec 2 [cited 2021 Dec 9];151(2):125–40. Available from: https://pubmed.ncbi.nlm.nih.gov/21962867/

52. Miller WR, Munita JM, Arias CA. Mechanisms of antibiotic resistance in enterococci. Expert review of anti-infective therapy [Internet]. 2014 Oct 1 [cited 2021 Dec 9];12(10):1221–36. Available from: https://pubmed.ncbi.nlm.nih.gov/25199988/

53. Patel R. Enterococcal-type glycopeptide resistance genes in non-enterococcal organisms. FEMS microbiology letters [Internet]. 2000 Apr [cited 2021 Dec 9];185(1):1–7. Available from: https://pubmed.ncbi.nlm.nih.gov/10731599/

54. Vescovo M, Morelli L, Bottazzi V. Drug resistance plasmids in Lactobacillus acidophilus and Lactobacillus reuteri. Applied and environmental microbiology [Internet]. 1982 [cited 2021 Dec 9];43(1):50–6. Available from: https://pubmed.ncbi.nlm.nih.gov/6798933/

55. Jarocki P, Podlešny M, Glibowski P, Targoński Z. A new insight into the physiological role of bile salt hydrolase among intestinal bacteria from the genus Bifidobacterium. PloS one [Internet]. 2014 Dec 1 [cited 2021 Dec 29];9(12). Available from: https://pubmed.ncbi.nlm.nih.gov/25470405/

56. Castro-Bravo N, Wells JM, Margolles A, Ruas-Madiedo P. Interactions of Surface Exopolysaccharides From Bifidobacterium and Lactobacillus Within the Intestinal Environment. Frontiers in microbiology [Internet]. 2018 Oct 11 [cited 2022 Jan 11];9(OCT). Available from: https://pubmed.ncbi.nlm.nih.gov/30364185/

57. Prasanna PHP, Grandison AS, Charalampopoulos D. Bifidobacteria in milk products: An overview of physiological and biochemical properties, exopolysaccharide production, selection criteria of milk products and health benefits. Food Research International. 2014 Jan 1;55:247–62.

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 of the United States of America [Internet]. 2012 Feb 7 [cited 2022 Jan 11];109(6):2108–13. Available from: /pmc/articles/PMC3277520/

59. Provencher C, LaPointe G, Sirois S, van Calsteren MR, Roy D. Consensus-degenerate hybrid oligonucleotide primers for amplification of priming glycosyltransferase genes of the exopolysaccharide locus in strains of the Lactobacillus casei group. Applied and environmental microbiology [Internet]. 2003 Jun 1 [cited 2022 Jan 11];69(6):3299–307. Available from: https://pubmed.ncbi.nlm.nih.gov/12788729/

60. 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 [Internet]. 2016 Apr 1 [cited 2022 Jan 11];92(4):1–12. Available from: https://pubmed.ncbi.nlm.nih.gov/26960391/

61. Yan S, Zhao G, Liu X, Zhao J, Zhang H, Chen W. Production of exopolysaccharide by Bifidobacterium longum isolated from elderly and infant feces and analysis of priming glycosyltransferase genes. RSC Advances [Internet]. 2017 Jun 16 [cited 2022 Jan 11];7(50):31736–44. Available from: https://pubs.rsc.org/en/content/articlehtml/2017/ra/c7ra03925e

Chapter 9

Gut microbe metabolism of small molecules supports human development across the early stages of life

Chiara Tarracchini, Federico Fontana, Leonardo Mancabelli, Gabriele Andrea Lugli, Giulia Alessandri, Francesca Turroni, Marco Ventura, Christian Milani

The results of this chapter were published in Frontiers in Microbiology, 2022 Sep; doi: 10.3389/fmicb.2022.1006721.

Abstract

From birth to adulthood, the human gut-associated microbial communities experience profound changes in their structure. However, while the taxonomical composition has been extensively explored, temporal shifts in the microbial metabolic functionalities related to the metabolism of bioactive small molecules are still largely unexplored.

Here, we collected a total of 6617 publicly available human fecal shotgun metagenomes and 42 metatranscriptomes from infants and adults to explore the dynamic changes of the microbial-derived small molecule metabolisms according to the age-related development of the human gut microbiome. Moreover, by selecting metagenomic data from 250 breastfed and 217 formula-fed infants, we also investigated how feeding types can shape the metabolic functionality of the incipient gut microbiome.

From the small molecule metabolism perspective, our findings suggested that the human gut microbial communities are genetically equipped and prepared to metabolically evolve toward the adult state as early as one month after birth, although at the age of four years, it still appeared functionally underdeveloped compared to adults. Furthermore, in respect of formula-fed newborns, breastfed infants showed enrichment in microbial metabolic functions related to specific amino acids present at low concentrations in human milk, highlighting that the infant gut microbiome has specifically evolved to synthesize bioactive molecules that can complement the human breast milk composition contributing to complete nutritional supply of infant.

For Supplementary Materials see the article published in Frontiers in Microbiology

INTRODUCTION

The complex microbial community associated with the human gut encompasses trillions of bacteria collectively referred to as the gut microbiota (1). The process of gut microbiota establishment is reported to be completed in a time window of approximately three years after childbirth (2). During this period, functional capabilities of the infant gut microbiome shift from the early lactate utilization towards plant polysaccharide breakdown, vitamin biosynthesis, and xenobiotic degradation, ultimately attaining adult-like microbiome capabilities (3,4). However, host and environmental factors, such as dietary habits, illness, and antibiotic treatments, continue to impact and modulate the gut microbiota structure from early infancy to adulthood (5).

As a large part of the microbial cells in the human gut are metabolically active, they are constantly influencing local and systemic host physiology. This ability is mediated by the production of thousands of unique bioactive small molecules, i.e., chemical compounds with a molecular weight < 3000 Da (6), which can accumulate in the intestine or reach organs and tissues through the blood circulatory system (7–9). These microbial metabolites can originate both from modifications to host-derived molecules, resulting in the production of branched- and short-chain fatty acids, secondary bile acids, and amino acids derivates such as tryptophan metabolites (10–14), or from *de novo* synthesis through secondary microbial metabolism (also known as specialized metabolism), which produce a wide range of molecules such as polyketides, nonribosomal peptides (NRPs), terpenes, NRP synthetase–independent siderophores, and saccharides (15,16). Accordingly, a large assortment of small molecules has been isolated from human gut-associated bacteria, highlighting their close involvement in host cellular functions and disease (17–19). However, the small molecule repertoire from the human gut microbiota and its evolution from infancy to adulthood have been poorly explored. In this context, we

exploited 6617 publicly available shotgun metagenomic and 42 metatranscriptomic sequencing data of fecal samples from infants (aged 0-4 years) and adults to infer microbial metabolic features related to small molecules and secondary metabolites production across different stages of human life.

RESULTS AND DISCUSSION

Collection of a very comprehensive metagenomic shotgun dataset

A total of 6,617 publicly available shotgun metagenomic samples from 4,062 healthy infants aged between a few days of life to four years of life were collected and clustered according to age (Table S1). Specifically, similar to what has been performed previously (20), infant age groups were named A (n = 732, 0-1 month of age), B (n = 1,209, 1-6 months of age), C (n = 788, 6-12 months of age), D (n = 922, 12-24 months of age), and E (n = 411, 24-48 months of age). In order to inspect the impact of the feeding type on the developing gut microbiome, from age groups A and B, we selected 250 metagenomic fecal samples from breastfed infants and 217 from newborns fed with infant formula based on the accessible metadata associated with the published studies (Table S1). Additionally, to generate a comprehensive dataset, a total of 2,555 shotgun metagenomic fecal samples from 18-70 years old human adults were retrieved from public repositories and assigned to the age group F (Table S1).

Moreover, in order to inspect the gene expression of microbial small molecule (mSM)-related functions throughout the maturation of the human gut microbiota, we collected a dataset of 42 public metatranscriptomic samples from infants (n = 26, 1-6 months of age) and adults (n = 16).

Species-level taxonomic profiles of microbial communities across age

The collected metagenomic shotgun sequencing data were used to longitudinally assess the relative abundance of individual gut-associated microbial species from infancy to adulthood (Table S2a,b). As fairly well established, the complexity and phylogenetic diversity of the infant gut microbiota progressively increase over time (Species richness, p-value < 0.005; Table S2a,b) while undergoing gradual changes

in community composition toward the adult-like microbiota (PERMANOVA *p*-value of < 0.001; Figure S1a). In particular, the milk-based diet is associated with the presence of *Escherichia coli* (average relative abundance of 9.08 % at 1-6 months) and members of the *Bifidobacterium* genus such as *B. longum*, *B. bifidum*, and *B. breve* (average relative abundance of 14.64 %, 6.86 %, and 9.10 %, respectively, at 1-6 months; Figure S1b, Table S2a), whose abundances tend to decrease following the weaning phase (4.62 %, 9.94 %, 5.04 %, and 5.74 %, respectively, around one year after birth, Table S2a).

In contrast, microbial species typical of the adult gut microbiota, such as *Eubacterium rectale*, *Bacteroides uniformis*, *Bacteroides fragilis*, *Prevotella copri*, and *Faecalibacterium prausnitzii*, begin to appear with the introduction of solid foods and changes in milk consumption, reaching the highest abundance in the adulthood (Figure S1b, Table S2a,b). Interestingly, fecal metagenomic data collected between 6 to 24 months after birth showed the greatest inter-individual variability among infants (C and D age groups, Figure S1a), supporting the assumption that the weaning phase profoundly impacts the gastrointestinal environment, leading to substantial developmental adaptations of the gut microbial communities (21).

Development of fecal microbial metabolic functionalities with age

As mentioned earlier, consequently to the microbial metabolic activities, thousands of bioactive mSM can be produced at the host-microbiota interface, shaping both local and systemic host physiology and eventually influencing human health (15). In this context, the shotgun metagenomic data were used to explore the evolution of the potential functionalities related to mSM biosynthesis throughout the infant gut microbiome maturation. For this purpose, we classified the metagenomic sequenced reads according to the MetaCyc database, revealing age-associated macro-

differences in the mSM metabolic profiles (PERMANOVA p-value < 0.05, Figure 1a, Table S3).

Specifically, following correction for multiple comparisons (Bonferroni Post Hoc test p-value < 0.05), a total of 271 unique microbial metabolic reactions codified through Enzyme Commission (EC) numbers were identified as statistically different between the age groups (Table S3).

Notably, among the microbial metabolic activities almost exclusively present in the first year after birth (age groups A, B, and C), we found enzymes for the metabolism of milk-derived compounds, such as D-tagatose (monosaccharide; EC number 4.1.2.40), ethanolamine (amino alcohol; EC number 4.3.1.7), N-Acteylglucosamine and Lacto-N-biose (N-glycans; EC numbers 3.2.1.52 and 3.2.1.140) (Figure 1b, Table S3). On average, their abundances progressively decreased from 0.011 % in age group A to 0.0053 % in age group C, in accordance with the presence of a one-year-lasting milk-adapted microbiota (22–24).

Otherwise, microbial functions involved in amino acids metabolism, including L-threonine, L-isoleucine, L-valine, and L-methionine biosynthesis (EC numbers 4.2.3.1, 4.2.1.9, and 2.5.1.48), as well as L-tryptophan catabolism with indole biosynthesis (EC number 4.2.1.20) were highly represented in the growing infants (mainly from a few days to one year of age) while maintaining their relevance also in the adult population (Figure 1b, Table S3). Interestingly, this may imply that the biosynthesis of crucial microbial-derived substrates to nutrients, such as essential amino acids, is ensured at all stages of life despite the microbiome compositional changes due to age-related diet diversification (25).

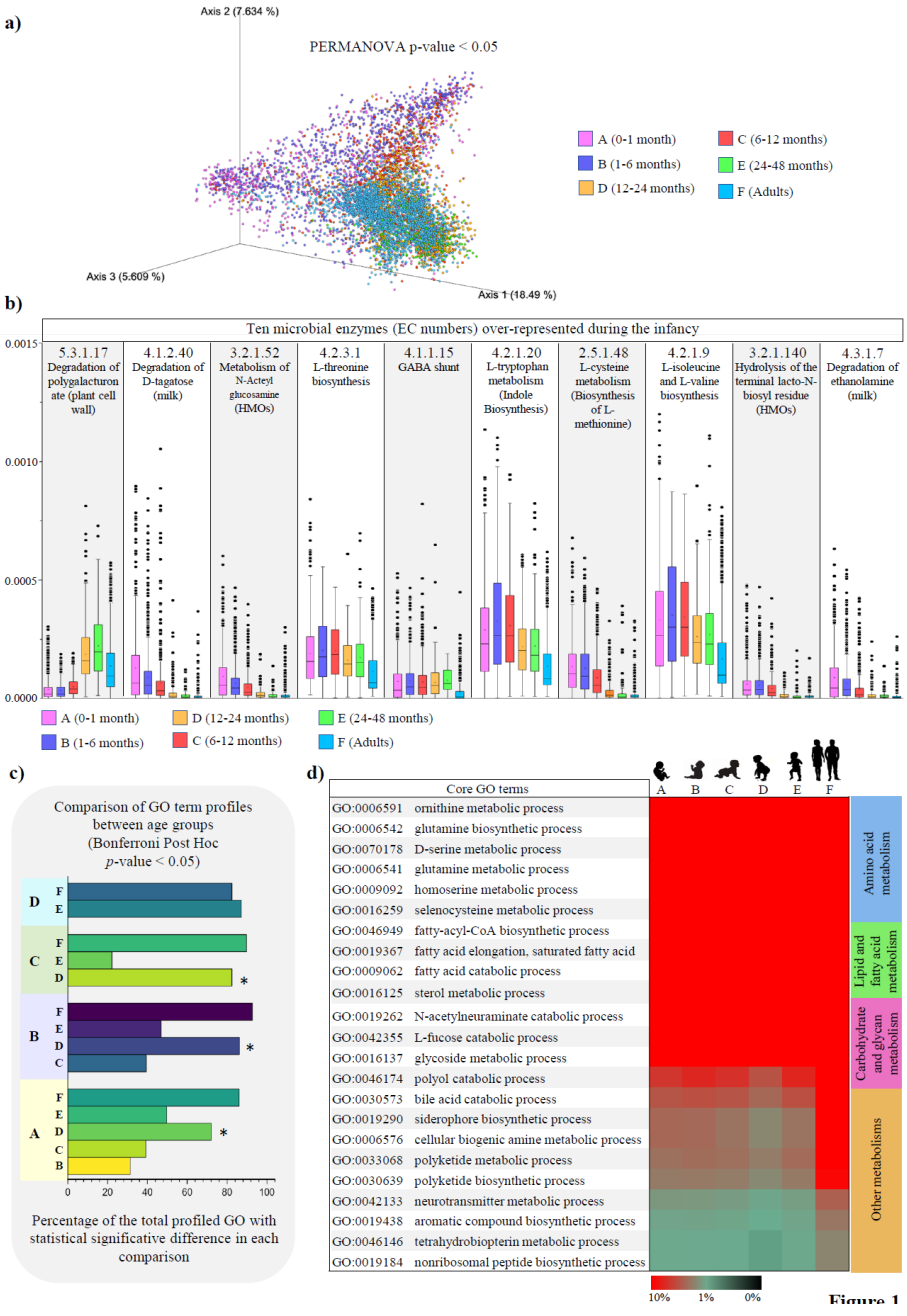

a)
Axis 2 (7.634 %)
PERMANOVA p-value < 0.05
Axis 3 (5.609 %)
Axis 1 (18.49 %)
A (0-1 month)
B (1-6 months)
C (6-12 months)
D (12-24 months)
E (24-48 months)
F (Adults)
b)
Ten microbial enzymes (EC numbers) over-represented during the infancy
0.0015
0.0010
0.0005
0.0000
5.3.1.17 Degradation of polygalacturonate (plant cell wall)
4.1.2.40 Degradation of D-tagatose (milk)
3.2.1.52 Metabolism of N-Acteyl glucosamine (HMOs)
4.2.3.1 L-threonine biosynthesis
4.1.1.15 GABA shunt
4.2.1.20 L-tryptophan metabolism (Indole Biosynthesis)
2.5.1.48 L-cysteine metabolism (Biosynthesis of L-methionine)
4.2.1.9 L-isoleucine and L-valine biosynthesis
3.2.1.140 Hydrolysis of the terminal lacto-N-biosyl residue (HMOs)
4.3.1.7 Degradation of ethanolamine (milk)
A (0-1 month)
B (1-6 months)
D (12-24 months)
C (6-12 months)
E (24-48 months)
F (Adults)
c)
Comparison of GO term profiles between age groups (Bonferroni Post Hoc p-value < 0.05)
D
C
B
A
0
20
40
60
80
100
Percentage of the total profiled GO with statistical significative difference in each comparison
d)
Core GO terms
A
B
C
D
E
F
GO:0006591 ornithine metabolic process
GO:0006542 glutamine biosynthetic process
GO:0070178 D-serine metabolic process
GO:0006541 glutamine metabolic process
GO:0009092 homoserine metabolic process
GO:0016259 selenocysteine metabolic process
GO:0046949 fatty-acyl-CoA biosynthetic process
GO:0019367 fatty acid elongation, saturated fatty acid
GO:0009062 fatty acid catabolic process
GO:0016125 sterol metabolic process
GO:0019262 N-acetylneuraminate catabolic process
GO:0042355 L-fucose catabolic process
GO:0016137 glycoside metabolic process
GO:0046174 polyol catabolic process
GO:0030573 bile acid catabolic process
GO:0019290 siderophore biosynthetic process
GO:0006576 cellular biogenic amine metabolic process
GO:0033068 polyketide metabolic process
GO:0030639 polyketide biosynthetic process
GO:0042133 neurotransmitter metabolic process
GO:0019438 aromatic compound biosynthetic process
GO:0046146 tetrahydrobiopterin metabolic process
GO:0019184 nonribosomal peptide biosynthetic process
Amino acid metabolism
Lipid and fatty acid metabolism
Carbohydrate and glycan metabolism
Other metabolisms
10%
1%
0%

Figure 1

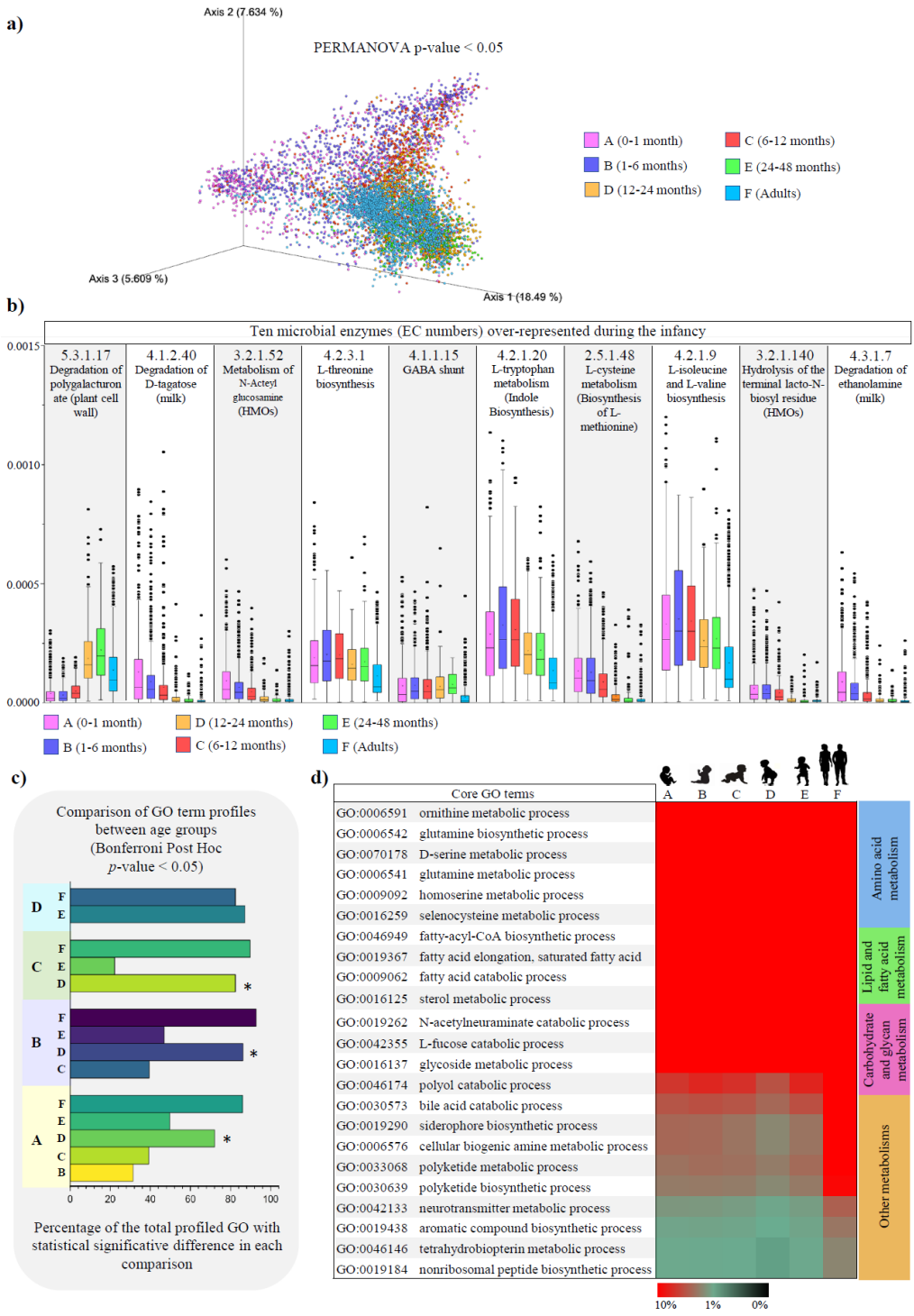


Figure 1. Differences in microbial pathways related to small molecule metabolisms from infancy to adulthood. Panel **(a)** shows the PCoA describing the age-associated differences (Bray-Curtis distances) in microbial gene profiles involved in small molecule metabolic pathways according to the MetaCyc database. Panel **(b)** reports unique metabolic reactions

(EC number) preferentially represented in infancy. Average abundances relative to the total profiled reads are reported on the vertical axis. Panel **(c)** showed the number of statistically different small molecule-associated biological processes observed from pairwise comparisons between age groups. Panel **(d)** depicts the microbial biological processes highly represented from infancy to adulthood. Heatmap colors represent the percentage of sequenced reads assigned to a specific microbial biological process (GO term).
Different colors indicate different age groups (0-1 month, pink; 1-6 months, violet; 6-12 months, red; 12-24 months, orange; 24-48 months, green; adults, light blue)

Gene Ontology classification of the enzymatic repertoire involved in mSM metabolism

In order to gain an overview of the age-associated variation in microbial biological processes, we classified the profiled microbial enzyme-encoding genes according to the Gene Ontology (GO) system (Table S4) (26).

Notably, by comparing the obtained microbial functional repertoires between age groups, we observed that age group D (12-24 months of age) showed the highest number of statistically significant GO terms in each comparison (ANOVA, p-value < 0.05; Figure 1c), thus diverging from the patterns observed in any other infancy time-point (Figure 1c). This may describe the dramatic changes in the microbial community composition undertaken concurrently with the cessation of milk intake. Indeed, it has been argued that the conclusion of milk-based diet (breast milk and /or formula) rather than the introduction of solid food at around six months of age induces profound modifications in the infant gut microbiome structure, leading it toward an adult-like state (27).

However, among the first 20 more abundant mSM core metabolism, i.e., microbial biological processes (GO terms) highly represented from birth to adulthood, we identified functions related to the production and degradation of fatty acids, biogenic amines (many of which act as eukaryotic neurotransmitters), glutamine, polyketides and nonribosomal peptides (which exhibit narrow-spectrum antimicrobial activities),

aromatic amino acids (L-tryptophan, L-tyrosin, and L-phenylalanine), and siderophores, along with specific functions for the catabolism of bile acids, L-fucose, and N-acetylneuraminate (sialic acid) (Figure 1d). Remarkably, these activities appear crucial for intestinal niche colonization by mediating bacterial competition, quorum sensing, and the utilization of the available carbohydrate sources, including human milk oligosaccharides and mucin (15,28–32).

Furthermore, as exemplified in Figure 2a reporting only the first 100 more abundant microbial-derived GO terms of biological processes, nearly all (84 %) of the statistically significant mSM-associated microbial activities profiled in adulthood (minimum abundance 0.05 %) were already present in one-month-old infants, albeit with a significantly lower abundance (minimum abundance 0.03 %, ANOVA p-value < 0.05) (Table S4). Thus, highlighting an overall progressive enhancement of the early-established microbial metabolic potential over time.

Specifically, synthesis or utilization of (branched-chain, sulfur, and aromatic) amino acids, biogenic amines such as putrescine, and vitamins appeared one month after birth with abundances ranging from 0.02 % to 0.07 % and then significantly enriched in adulthood (abundances between 0.05 % and 0.19 %) (Bonferroni Post Hoc test p-value < 0.05) (Figure 2b). In contrast, synthesis pathways of butyric acid, terpenoids, and nonprotein amino acids such as citrulline, along with metabolisms of cholesterol, mannose, and xylose, were found among the small molecule metabolisms nearly uniquely present in adulthood (Figure 2b).

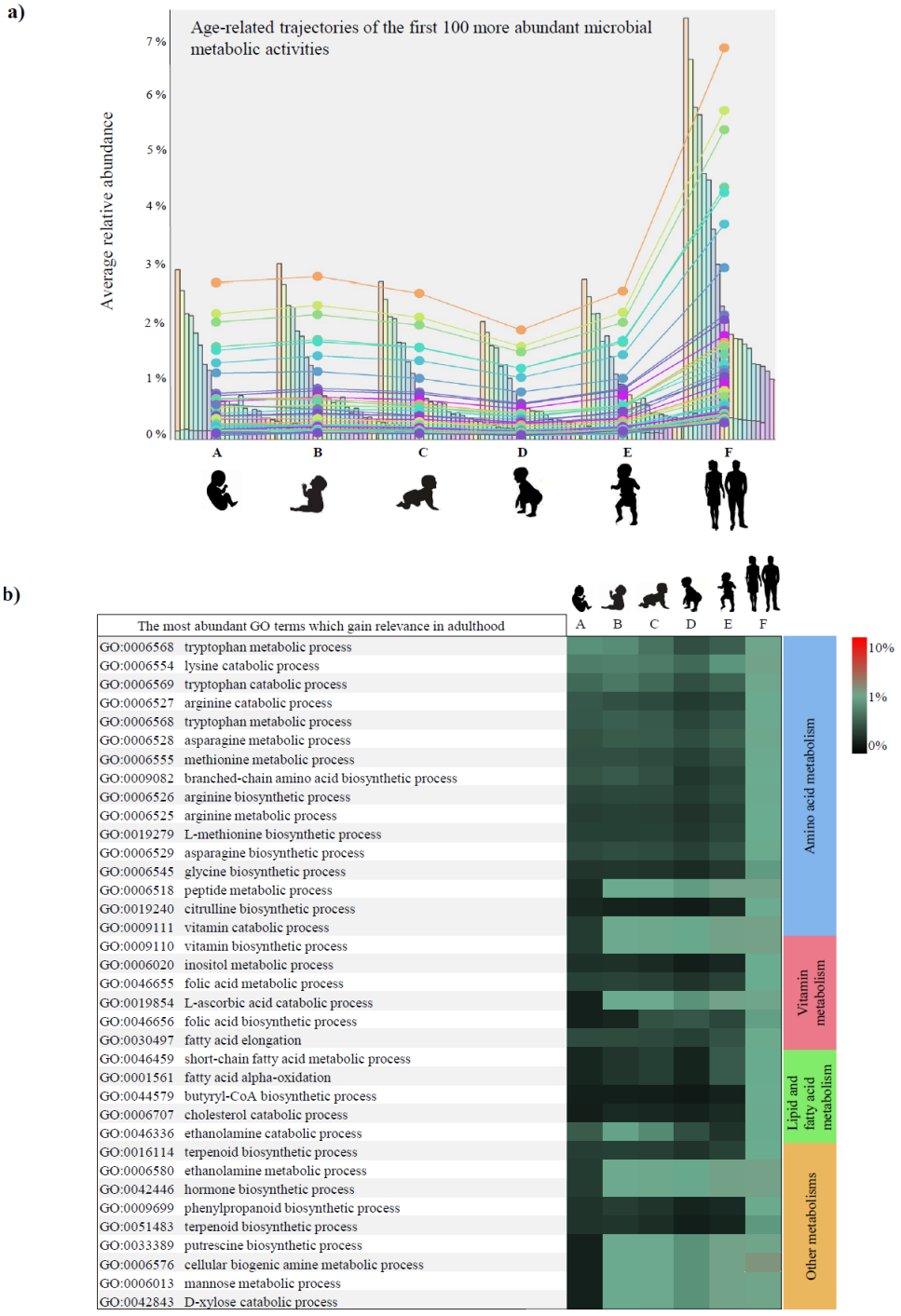


Figure 2. Developmental trajectories of microbial small-molecule-related functions. Panel **(a)** shows the developmental trend in the relative abundance of the small molecule-related microbial functions. In panel **(b),** relative abundances of the leading small molecule metabolisms enriched in adulthood rather than infancy are illustrated by a heatmap. The

heatmap color scale illustrates the percentage of sequenced reads assigned to a specific microbial biological process (GO term).
Different letters indicates different age groups (0-1 month, A ; 1-6 months, B; 6-12 months, C; 12-24 months, D; 24-48 months, E; adults, F)

All the microbial functions found preferentially associated with adulthood are believed to provide several metabolites relevant for the microbe-host mutualisms, contributing positively to the host physiology. In particular, butyric acid and polyamines, including putrescine, have demonstrated beneficial effects on the human gut mucosa (33), while vitamins and various amino acids can translocate in systemic circulation and exert far-reaching effects on the host (34,35). Therefore, the infant gut microbiota maturation emerged to be far from concluded even at 3-4 years of age, albeit the taxonomic composition may resemble that of adults (36).

However, from about one month after birth, infant gut-associated microbial communities appeared genetically equipped with most of the microbial metabolic functions that will support intestinal homeostasis and physiological processes in adults, suggesting a very early foundation of the host-microbiota symbiosis, with improvements in the microbial metabolic potential throughout the host development. These data indicate that host-microbiome co-evolution led to the selection of microbial genetic traits necessary for survival and growth in the varying intestinal niche as well as for the perpetual production of small molecules of high biological relevance in the host physiology.

Early infant feeding practices shape the metabolic traits of the fecal microbial communities

It is very well known that the feeding type is among the key factors influencing the infant gut microbiota composition and, therefore, gastrointestinal functions (37). In order to evaluate the gut microbiome structure as a function of the infant diet

(breastfeeding vs. formula feeding), we considered 250 shotgun metagenomic samples from breastfed newborns and 217 from formula-fed infants from the age group A (0-1 month) and B (1-6 months) (Table S1). As highlighted in Figure S2a, except for the 0-1 month-old infants showing high inter-individual variability, infant diet is associated with distinct gut microbial compositional patterns (PERMANOVA p-value < 0.05; Figure S2a), with significantly lower growth of *E. coli* throughout the lactation and higher abundance of *B. breve* and *B. bifidum* at 1-6 months in breastfed infants rather than in those fed with infant formula (t-test p-value < 0.05) (Figure S2b).

Classification of the enzyme-encoding genes using the MetaCyc database of small molecules, combined with functional enrichment based on the GO annotation system, revealed microbial functional diversity according to the feeding type (PERMANOVA p-value < 0.05; Figure 3a).

In particular, compared with formula-fed infants, those receiving breast milk were enriched in microbial metabolisms involved in the degradation of L-fucose (GO:0042354), (GO:0030573), and N-acetylneuraminate (GO:0006054), along with biosynthetic pathways of indole-3-acetic acid (L-tryptophan metabolism, GO:0006569), and polyketides (GO:0030639) (Figure 3b, c; Table S5, S6). Moreover, microbial production of vitamins, including folic acid (vitamin B9, GO:0046656) and biotin (vitamin B7; EC number 6.3.4.15), became significantly preponderant in breastfed infants 1-6 months after birth (Figure 3b; Table S5). This could explain why, although human breast milk contains a slightly low concentration of biotin, no signs of biotin deficiency were noted in breastfed newborns (37).

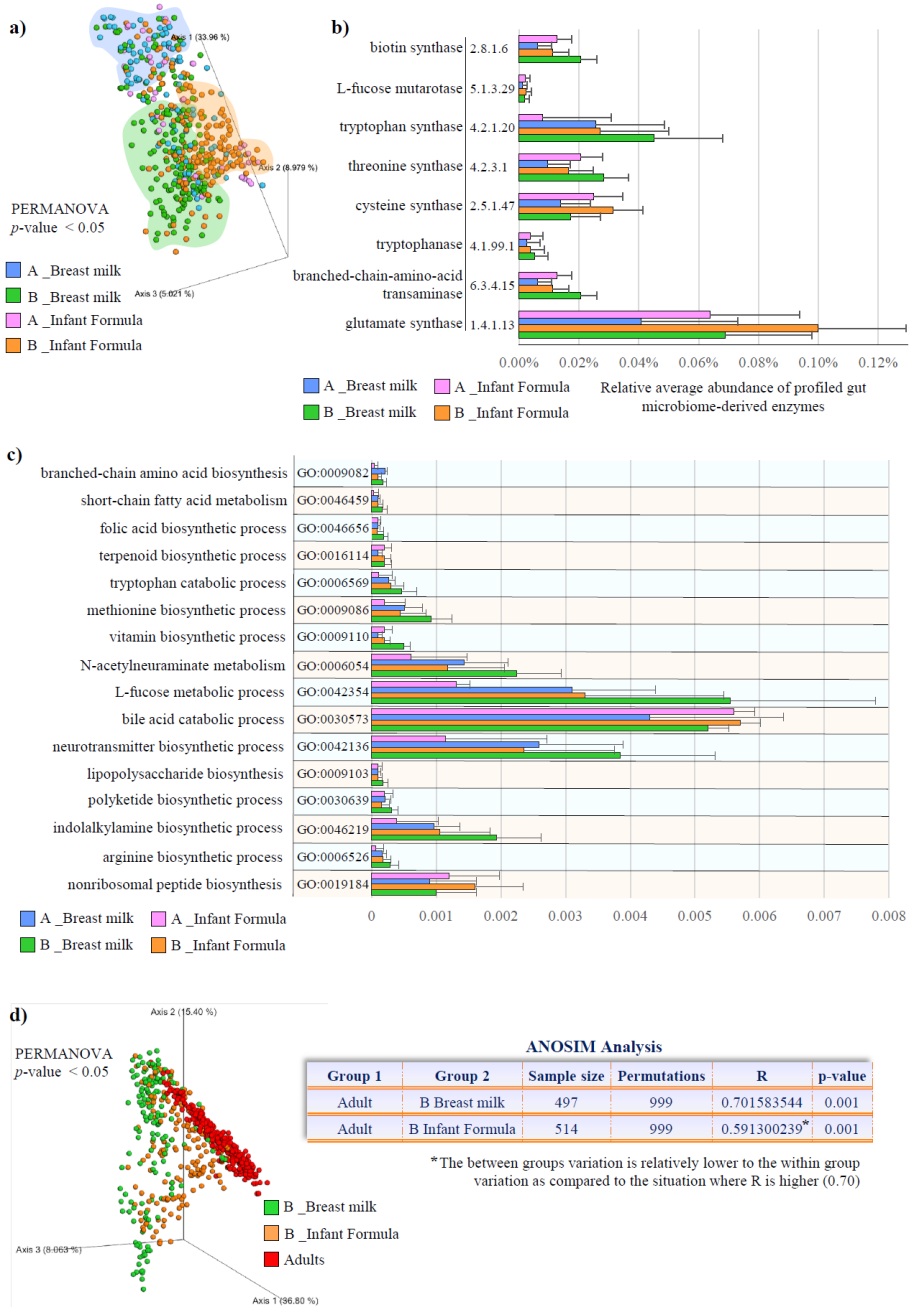


Figure 3. Comparison of the microbial small molecule profiles between metagenomes of breastfed and formula-fed infants. In panel **(a)**, the PCoA depicts the significant differences in the small molecule profiles between breast-fed and formula-fed infants aged 0-1 months (A) and 1-6 months (B). Panel **(b)** exhibits the microbial small molecule-related pathways (EC number) with statistically significant differences between feeding types. Panel **(c)** shows differentially abundant microbial biological processes involved in small molecule metabolisms classified according to the Gene Ontology (GO) annotation system. In panel

(d), PCoA and ANOSIM analysis compare the small molecule repertoires observed in different feeding practices (breastfeeding and formula) and adults.
Different colors indicate different ages and feeding types (0–1-month-old breastfed, light blue; 1-6 months old breastfed, green; 0–1-month-old formula-fed, pink; 1-6 months old formula fed, orange).

Similarly, breastfed infants showed an increase in microbial biological processes involved in the biosynthesis of L-arginine (GO:0006526), L-methionine (GO:0009086), L-tryptophan (EC number 4.2.1.20), L-threonine (EC number 4.2.3.1), L-cysteine (EC number 2.5.1.47), L-glutamate (EC number 1.4.1.13) and branched-chain amino acids (GO:0009082) (Bonferroni Post Hoc test p-value < 0.05) (Figure 3b,c, Table S6), corresponding to the amino acids with a lower concentration in breastmilk (38). More specifically, while the human milk content of these latter amino acids declines with lactation progression (39), we found that their microbial production tends to increase, completing breast milk composition throughout the natural dynamic changes of lactation, thus supporting the high protein requirement in developing infants (40,41).

Beyond the specific microbial activities that we found enriched in breastfed infants, microbial functional analyses highlighted that most (64%) of the profiled GO terms were more abundant in formula-fed infants, indicating that compared with breastfed newborns, the gut microbiome of infants receiving formula resembles earlier that of adults (ANOSIM R=0.59 vs. R=0.70 at p-value=0.001; Figure 3d, Table S6). Accordingly, this evidence suggested that formula-fed infant gut microbiota may evolve more rapidly compared with that of their breastfed counterpart, as previously partially observed solely from a taxonomic point of view through the use of specific metrics, i.e., “relative microbiota maturity” and “microbiota-for-age Z score” (42,43). Thus, lack of breast milk intake could preclude the gradual specialization of the gut microbiota that instead parallelly accompanies the growth of the breastfed

newborn with the well-known positive effects on intestinal, neurological, and immune system development (44–46).

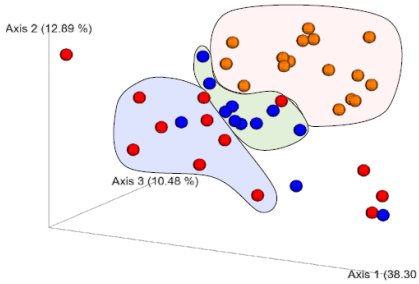
Altogether, these findings widen the repertoire of the notorious benefits of breastfeeding to the production of bioactive mSM with relevant biological roles in the host.

Changes in the metatranscriptome profiles driving small molecule metabolisms along the infant gut microbiome development

In order to inspect expression patterns of microbial genes related to the mSM metabolism in the human gut microbiota from infancy to adulthood, we collected a total of 42 public available cross-sectional metatranscriptomic samples from infants aged between one month and one year along with adults (18-70 years) (Table S1). According to age, the datasets were then gathered in age groups TR_B (n = 14, 1-6 months old), TR_C (n = 12, 6-12 months old), and TR_F (n = 16, adults).

Similarity analysis (ANOSIM) and visualization of the variation dispersal at the metatranscriptome level (PCoA) showed that differences among samples within infant age groups TR_B and TR_C (including those resulting from different delivery routes and feeding types) were lower than those emerging by comparison with TR_F, suggesting that age-related factors are the main forces driving the maturation of the gut metatranscriptome (Figure 4a).

a)

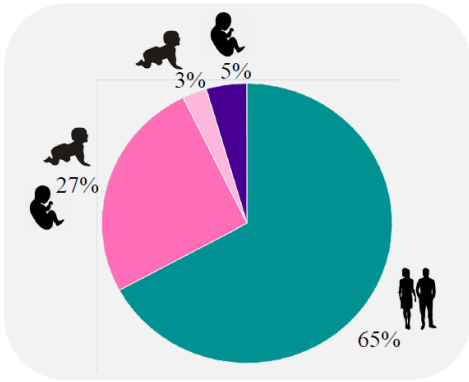


ANOSIM Analysis

Group 1	Group 2	Sample size	Permutations	R	p-value	q-value
TR_B	TR_C	26	999	0.04788	0.112	0.112
TR_B	TR_F	30	999	0.39017	**0.001**	0.0015
TR_C	TR_F	28	999	0.48538	**0.001**	0.0015

TR_F (Adults)
TR_C (6-12 months old infants)
TR_B (1-6 months old infants)

b)



c)

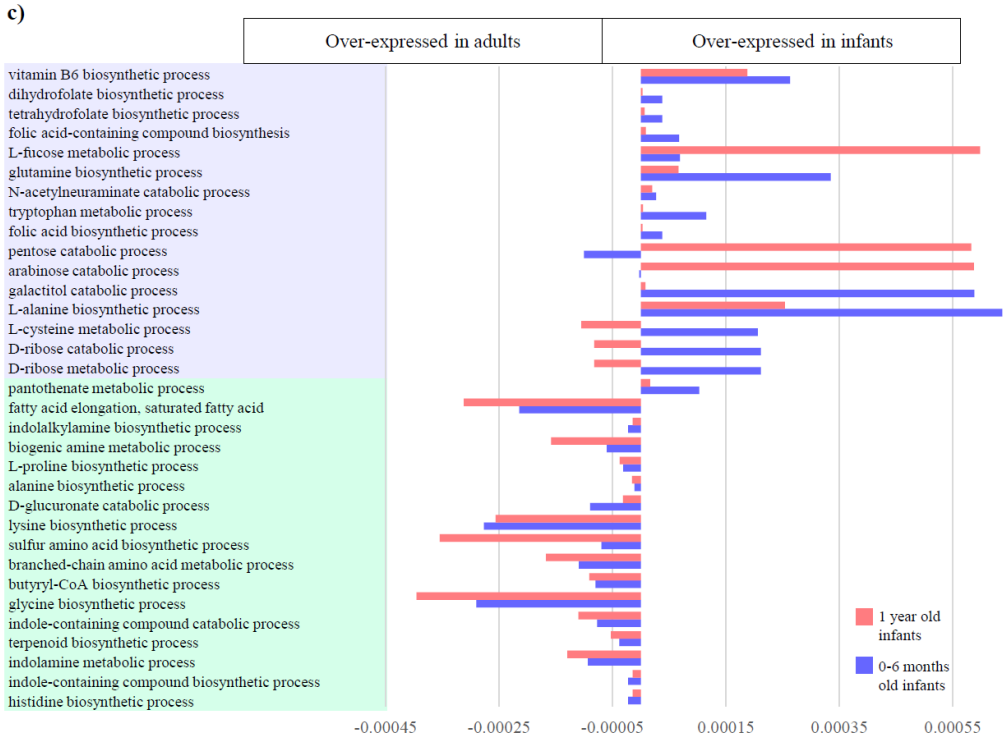


Figure 4. Differences in the expression of microbial functions related to small molecule metabolism based on metatranscriptome level analysis. In panel **(a)**, macro differences in the small molecule gene expression patterns between infants and adults are described through PCoA and ANOSIM analysis. Different colors indicate different ages groups (1-6 months

old, TR_B, red; 6-12 months old, TR_C, blue; and adults, TR_F, orange). In panel **(b)**, pie chart depicts the distribution among infant and adult metatranscriptomes of the differentially expressed genes encoding small molecule-related functions. Different colors indicate genes overexpressed in different age groups (genes overexpressed in adults, green; genes overexpressed in 0-6 months old OR in 1-year-old OR in both infants ages, dark pink; genes overexpressed in 0-6 months old infants, light pink; genes overexpressed in 1-year-old infants, violet). Panel **(c)** reports developmental changes in the expression of selected small molecule-related metabolisms.

Overall, the metatranscriptome-level analysis revealed that 67 % of mSM-related microbial activities were overexpressed in adults, implying that the reduced genetic potential for microbial functionality characterizing the infant gut microbiome, described above compared to adults, is also reflected at the metatranscriptome level (Figure 4b; Table S7).

Specifically, adult-like metatranscriptomes showed increased expression of microbial biological processes involved in a wide range of amino acid metabolisms, including branched-chain and sulfur-containing amino acids, as well as L-lysine, L-histidine, and L-proline (Figure 4c; Table S7). Besides, also functions related to the metabolism of fatty acids, biogenic amines, terpenoids, and indole-containing compounds, as well as butyric acid biosynthesis, were significantly over-expressed in adulthood (Figure 4c; Table S7), thus, evidencing a wide variety of mSM related activities with a progressive evolution of their genes' expression patterns over time.
In contrast, infants showed the over-expression of biological processes involved in the catabolism of L-tryptophan, along with biosynthetic pathways of several vitamins, including B9, B5, B6, and K2, L-glutamine, L-methionine, L-cysteine, and L-homoserine, which acts as a precursor for methionine, threonine, and isoleucine (Figure 4c, Table S7). Also, degradation of monosaccharides, such as D-ribose, D-arabinose, D-glucose, D-galactose, and pentoses, along with pathways for the utilization of L-fucose and N-Acetylneuraminic acid, were overexpressed in infancy

compared to adulthood, suggesting overall an active biosynthetic metabolism that relies more on simple diet-derived sugars and host-derived oligosaccharides (Figure 4c, Table S7).

Interestingly, 65 % of the mSM-related microbial metabolisms over-expressed in infancy did not show significant abundance differences between infants and adults at the metagenome/gene level, while the remaining 35 % were among those we found less abundant in the corresponding infant metagenomes compared to adults. Altogether, these findings confirmed that developmental stage-specific mSM-associated microbial functions are accomplished through the accommodation of gene expression of the existing genetic metabolic potential rather than solely through rearrangements of the microbial genetic makeup.

Conclusion

Under homeostatic conditions, the gut microbiota metabolizes the diet- and host-derived substrates available in the intestinal environment, producing a multitude of bioactive metabolites, such as small molecules, with positive impacts on host physiology (17). However, developmental adjustments in microbial activities related to the small molecule metabolism in correlation with age remain poorly investigated. In this context, analyses of 6617 shotgun metagenome fecal samples from infants and adults highlighted that 4-years old infants still showed underdeveloped microbial activities compared to adults. This implies that although the infant gut microbiota is considered compositionally mature around three years after birth, developmental processes of small molecule-associated functionalities are still far from accomplished. However, as highlighted from the metatranscriptome perspective, most microbial genetic features for small molecule metabolisms that are highly expressed in adulthood were already present at one month of life, although with a lower abundance. Similarly, a portion of the small molecule-related functions

overexpressed in infancy corresponded to those with a lower abundance compared to adult metagenomes. This fact points to a fine modulation of the microbial small molecule-related activities through gene expression patterns that contribute to meeting stage-specific host physiological demands.

Furthermore, comparison of different infant feeding practices (breastfeeding vs. infant formula) showed that breastfed infants were enriched in microbial pathways involved in the biosynthesis of amino acids with low concentration in human milk, pointing out the existence of complementarity between breast milk and gut microbiome functionalities.

Altogether, these data highlighted that the close cooperation between gut bacteria and host cells relies on the early establishment of the microbial metabolic (genetic) potential, whose expression will be dynamically adapted in response to the availability of nutritive resources as well as the physiological needs of the specific host's developmental stages. In this regard, findings collected in this meta-analysis will be pivotal for future studies aimed at comparing infant small molecule profiles in different nutritional conditions as well as identifying the major gut commensals producers of such bioactive metabolites.

METERIALS AND METHODS

Sample collection

A total of 6617 publicly available shotgun metagenomic sequencing data of fecal samples from 4062 infants aged 0-4 years and 2555 adults (18-70 years) were collected from 63 different BioProjects of the Sequence Read Archive (SRA) database (Table S1). According to the associated metadata, infants were born at term via uncomplicated Cesarean or natural vaginal delivery and assumed or did not breast milk (Table S1). All subjects were overall healthy, had no history or clinical evidence

of any disease, and did not take antibiotics at the time of sample collection. Similarly, 42 metatranscriptomic data from the gut microbiome of 26 infants (aged between one month and one year) and 16 adults (18-70 years) were retrieved from the public repository of NCBI (Table S1).

Data processing of the metagenome and metatranscriptome sequences

After being recovered from the SRA public repositories, the collected shotgun sequencing data were filtered for quality (minimum mean quality score, 20; window size, 5 bp; and minimum length, 80 bp). Subsequently, metagenomics reads aligning/mapping to the *Homo sapiens* genome sequence were identified through blastn program and removed. Taxonomic profiling of the retained sequenced reads was achieved with the METAnnotatorX2 bioinformatics platform (47), using the up-to-date RefSeq (genome) sequence database retrieved from the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/refseq/). Species-level taxonomic classification of each read was achieved through Megablast (48) (with option -e-value 1e-5, -qcov_hsp_perc 50) using > 94% alignment identity. Reads that showed the same sequence identity against more than one bacterial species were discarded. Based on species abundance, the similarity between samples (beta-diversity) was computed using Bray-Curtis dissimilarity calculated for pairwise comparisons. Principal Coordinates Analysis (PCoA) representation of beta-diversity was performed using ORIGIN 2021 (https://www.originlab.com/2021).

Functional analyses based on MetaCyc database of small molecules and GO enrichment

An implemented function of the software METAnnotatorX2 (47) was employed for the functional classification of the metagenome- and metatranscriptome-derived microbial reads according to the MetaCyc database of small molecules (49).

Subsequently, for each of the profiled microbial enzymes involved in small molecule metabolism, we collected the associated GO terms (all parentals) through the Bioconductor annotation data package GO.db (release 3.15) (https://bioconductor.org/packages/GO.db/).

Statistical analyses

The software SPSS version 25, and ORIGIN version 9.8.0.200 (www.ibm.com/software/it/analytics/spss/) (https://www.originlab.com/) were used for statistical data analyses and graphing. Principal Coordinates Analysis (PCoA) based on Bray-Curtis dissimilarity matrix was performed using the software QIIME 2 (50). PERMANOVA analyses were conducted using 1000 permutations to estimate *p*-values for the observed differences between the compared groups in PCoA analyses. Similarity analysis (ANOSIM) was performed through QIIME 2 software with 999 random permutations on the same Bray-Curtis distance matrix obtained from the test for differences in small-molecule repertoires among the different groups.

AKNOWLEDGMENTS

This study was supported by “Fondi locali per la Ricerca 2020, Azione B - Progetti di ricerca riservati a giovani ricercatori” for the project entitled “Drawing the multi-omics atlas of infant gut microbiota”. We thank GenProbio Srl for the financial support of the Laboratory of Probiogenomics. Part of this research was conducted using the High-Performance Computing (HPC) facility of the University of Parma.

References

1. Thursby E, Juge N. Introduction to the human gut microbiota. Biochem J [Internet]. 2017 Jun 1 [cited 2022 Jul 27];474(11):1823–36. Available from: https://pubmed.ncbi.nlm.nih.gov/28512250/

2. Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO. Development of the human infant intestinal microbiota. PLoS Biol [Internet]. 2007 Jul [cited 2022 Jul 27];5(7):1556–73. Available from: https://pubmed.ncbi.nlm.nih.gov/17594176/

3. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, et al. Succession of microbial consortia in the developing infant gut microbiome. Proc Natl Acad Sci U S A [Internet]. 2011 Mar 15 [cited 2022 Jul 27];108 Suppl 1(Suppl 1):4578–85. Available from: https://pubmed.ncbi.nlm.nih.gov/20668239/

4. Avershina E, Lundgård K, Sekelja M, Dotterud C, Storrø O, Øien T, et al. Transition from infant- to adult-like gut microbiota. Environ Microbiol [Internet]. 2016 Jul 1 [cited 2022 Jul 27];18(7):2226–36. Available from: https://pubmed.ncbi.nlm.nih.gov/26913851/

5. Hasan N, Yang H. Factors affecting the composition of the gut microbiota, and its modulation. PeerJ [Internet]. 2019 [cited 2022 Jul 27];7(8). Available from: https://pubmed.ncbi.nlm.nih.gov/31440436/

6. Pauer H, Teixeira FL, Robinson A v., Parente TE, de Melo MAF, Lobo LA, et al. Bioactive small molecules produced by the human gut microbiome modulate Vibrio cholerae sessile and planktonic lifestyles. Gut Microbes [Internet]. 2021 [cited 2022 Jul 27];13(1):1–19. Available from: https://pubmed.ncbi.nlm.nih.gov/34006192/

7. Uchimura Y, Fuhrer T, Li H, Lawson MA, Zimmermann M, Yilmaz B, et al. Antibodies Set Boundaries Limiting Microbial Metabolite Penetration and the Resultant Mammalian Host Response. Immunity [Internet]. 2018 Sep 18 [cited 2022 Jul 27];49(3):545-559.e5. Available from: https://pubmed.ncbi.nlm.nih.gov/30193848/

8. Vojinovic D, Radjabzadeh D, Kurilshikov A, Amin N, Wijmenga C, Franke L, et al. Relationship between gut microbiota and circulating metabolites in population-based cohorts. Nat Commun [Internet]. 2019 Dec 1 [cited 2022 Jul 27];10(1). Available from: https://pubmed.ncbi.nlm.nih.gov/31862950/

9. Han J, Antunes LCM, Finlay BB, Borchers CH. Metabolomics: towards understanding host-microbe interactions. Future Microbiol [Internet]. 2010 Feb [cited 2022 Jul 27];5(2):153–61. Available from: https://pubmed.ncbi.nlm.nih.gov/20143941/

10. Morrison DJ, Preston T. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. Gut Microbes [Internet]. 2016 May 3 [cited 2022 Jul 27];7(3):189–200. Available from: https://pubmed.ncbi.nlm.nih.gov/26963409/

11. Dai ZL, Wu G, Zhu WY. Amino acid metabolism in intestinal bacteria: links between gut ecology and host health. Front Biosci (Landmark Ed) [Internet]. 2011 Jan 1 [cited 2022 Jul 27];16(5):1768–86. Available from: https://pubmed.ncbi.nlm.nih.gov/21196263/

12. Ridlon JM, Harris SC, Bhowmik S, Kang DJ, Hylemon PB. Consequences of bile salt biotransformations by intestinal bacteria. Gut Microbes [Internet]. 2016 Jan 2 [cited 2022 Jul 27];7(1):22–39. Available from: https://pubmed.ncbi.nlm.nih.gov/26939849/

13. Ridlon JM, Kang DJ, Hylemon PB, Bajaj JS. Bile acids and the gut microbiome. Curr Opin Gastroenterol [Internet]. 2014 [cited 2022 Jul 27];30(3):332–8. Available from: https://pubmed.ncbi.nlm.nih.gov/24625896/

14. Liu Y, Dai M. Trimethylamine N-Oxide Generated by the Gut Microbiota Is Associated with Vascular Inflammation: New Insights into Atherosclerosis. Mediators Inflamm [Internet]. 2020 [cited 2022 Jul 27];2020. Available from: https://pubmed.ncbi.nlm.nih.gov/32148438/

15. Donia MS, Fischbach MA. HUMAN MICROBIOTA. Small molecules from the human microbiota. Science [Internet]. 2015 Jul 24 [cited 2022 Jul 27];349(6246). Available from: https://pubmed.ncbi.nlm.nih.gov/26206939/

16. Postler TS, Ghosh S. Understanding the Holobiont: How Microbial Metabolites Affect Human Health and Shape the Immune System. Cell Metab [Internet]. 2017 Jul 5 [cited 2022 Jul 27];26(1):110–30. Available from: https://pubmed.ncbi.nlm.nih.gov/28625867/

17. Luber JM, Kostic AD. Gut Microbiota: Small Molecules Modulate Host Cellular Functions. Curr Biol [Internet]. 2017 Apr 24 [cited 2022 Jul 27];27(8):R307–10. Available from: https://pubmed.ncbi.nlm.nih.gov/28441565/

18. Tarracchini C, Milani C, Longhi G, Fontana F, Mancabelli L, Pintus R, et al. Unraveling the Microbiome of Necrotizing Enterocolitis: Insights in Novel Microbial and Metabolomic Biomarkers. Microbiol Spectr [Internet]. 2021 Oct 31 [cited 2022 Jul 27];9(2). Available from: https://pubmed.ncbi.nlm.nih.gov/34704805/

19. Bekkers M, Stojkovic B, Kaiko GE. Mining the Microbiome and Microbiota-Derived Molecules in Inflammatory Bowel Disease. Int J Mol Sci [Internet]. 2021 Oct 1 [cited 2022 Jul 27];22(20). Available from: https://pubmed.ncbi.nlm.nih.gov/34681902/

20. Mancabelli L, Tarracchini C, Milani C, Lugli GA, Fontana F, Turroni F, et al. Multi-population cohort meta-analysis of human intestinal microbiota in early life reveals the existence of infant community state types (ICSTs). Computational and Structural Biotechnology Journal [Internet]. 2020 Jan 1 [cited 2021 Sep 20];18:2480–93. Available from: https://pubmed.ncbi.nlm.nih.gov/33005310/

21. Dizzell S, Stearns JC, Li J, van Best N, Bervoets L, Mommers M, et al. Investigating colonization patterns of the infant gut microbiome during the introduction of solid food and weaning from breastmilk: A cohort study protocol. PLoS ONE [Internet]. 2021 Apr 1 [cited 2021 Nov 11];16(4 April). Available from: https://pubmed.ncbi.nlm.nih.gov/33798237/

22. Campbell HR, Alsharif FM, Marsac PJ, Lodder RA. The Development of a Novel Pharmaceutical Formulation of D-Tagatose for Spray-Drying. Journal of Pharmaceutical Innovation [Internet]. 2022 Mar 1 [cited 2022 Jul 27];17(1):194–206. Available from: https://link.springer.com/article/10.1007/s12247-020-09507-4

23. Zhou J, Xiong X, Wang KX, Zou LJ, Ji P, Yin YL. Ethanolamine enhances intestinal functions by altering gut microbiome and mucosal anti-stress capacity in weaned rats. Br J Nutr [Internet]. 2018 Aug 14 [cited 2022 Jul 27];120(3):241–9. Available from: https://pubmed.ncbi.nlm.nih.gov/29789027/

24. Tarracchini C, Milani C, Lugli GA, Mancabelli L, Fontana F, Alessandri G, et al. Phylogenomic disentangling of the Bifidobacterium longum subsp. infantis taxon. Microb Genom [Internet]. 2021 [cited 2022 Jul 27];7(7). Available from: https://pubmed.ncbi.nlm.nih.gov/34319225/

25. Odamaki T, Kato K, Sugahara H, Hashikura N, Takahashi S, Xiao JZ, et al. Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. BMC Microbiol [Internet]. 2016 May 25 [cited 2022 Jul 27];16(1). Available from: https://pubmed.ncbi.nlm.nih.gov/27220822/

26. Harris MA, Deegan JI, Ireland A, Lomax J, Ashburner M, Tweedie S, et al. The Gene Ontology project in 2008. Nucleic Acids Res [Internet]. 2008 Jan 1 [cited 2022 Jul 27];36(Database issue). Available from: https://pubmed.ncbi.nlm.nih.gov/17984083/

27. Davis EC, Wang M, Donovan SM. The role of early life nutrition in the establishment of gastrointestinal microbial composition and function. Gut Microbes [Internet]. 2017 Mar 4 [cited 2022 Jul 27];8(2):143–71. Available from: https://pubmed.ncbi.nlm.nih.gov/28068209/

28. Spano G, Russo P, Lonvaud-Funel A, Lucas P, Alexandre H, Grandvalet C, et al. Biogenic amines in fermented foods. Eur J Clin Nutr [Internet]. 2010 [cited 2022 Jul 27];64 Suppl 3:S95–100. Available from: https://pubmed.ncbi.nlm.nih.gov/21045859/

29. Pickard JM, Chervonsky A v. Intestinal fucose as a mediator of host-microbe symbiosis. J Immunol [Internet]. 2015 Jun 15 [cited 2022 Jul 27];194(12):5588–93. Available from: https://pubmed.ncbi.nlm.nih.gov/26048966/

30. Yu ZT, Chen C, Newburg DS. Utilization of major fucosylated and sialylated human milk oligosaccharides by isolated human gut microbes. Glycobiology [Internet]. 2013 Nov [cited 2022 Jul 27];23(11):1281–92. Available from: https://pubmed.ncbi.nlm.nih.gov/24013960/

31. Zimmermann M, Fischbach MA. A family of pyrazinone natural products from a conserved nonribosomal peptide synthetase in Staphylococcus aureus. Chem Biol [Internet]. 2010 Sep 24 [cited 2022 Jul 27];17(9):925–30. Available from: https://pubmed.ncbi.nlm.nih.gov/20851341/

32. Guzior D v., Quinn RA. Review: microbial transformations of human bile acids. Microbiome [Internet]. 2021 Dec 1 [cited 2022 Jul 27];9(1). Available from: https://pubmed.ncbi.nlm.nih.gov/34127070/

33. Tofalo R, Cocchi S, Suzzi G. Polyamines and Gut Microbiota. Front Nutr [Internet]. 2019 Feb 25 [cited 2022 Jul 27];6. Available from: https://pubmed.ncbi.nlm.nih.gov/30859104/

34. van der Wielen N, Moughan PJ, Mensink M. Amino Acid Absorption in the Large Intestine of Humans and Porcine Models. J Nutr [Internet]. 2017 Aug 1 [cited 2022 Jul 27];147(8):1493–8. Available from: https://pubmed.ncbi.nlm.nih.gov/28615378/

35. Magnúsdóttir S, Ravcheev D, de Crécy-Lagard V, Thiele I. Systematic genome assessment of B-vitamin biosynthesis suggests cooperation among gut microbes. Front Genet [Internet]. 2015 [cited 2022 Jul 27];6(MAR). Available from: https://pubmed.ncbi.nlm.nih.gov/25941533/

36. Zeng S, Ying J, Li S, Qu Y, Mu D, Wang S. First 1000 Days and Beyond After Birth: Gut Microbiota and Necrotizing Enterocolitis in Preterm Infants. Front Microbiol [Internet]. 2022 Jun 21 [cited 2022 Jul 27];13. Available from: https://pubmed.ncbi.nlm.nih.gov/35801107/

37. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature [Internet]. 2012 Jun 14 [cited 2022 Jul 27];486(7402):222–7. Available from: https://pubmed.ncbi.nlm.nih.gov/22699611/

38. Zhang Z, Adelman AS, Rai D, Boettcher J, Lonnerdal B. Amino acid profiles in term and preterm human milk through lactation: a systematic review. Nutrients [Internet]. 2013 Nov 26 [cited 2022 Jul 27];5(12):4800–21. Available from: https://pubmed.ncbi.nlm.nih.gov/24288022/

39. Zhang Z, Adelman AS, Rai D, Boettcher J, Lonnerdal B. Amino acid profiles in term and preterm human milk through lactation: a systematic review. Nutrients [Internet]. 2013 Nov 26 [cited 2022 Jul 27];5(12):4800–21. Available from: https://pubmed.ncbi.nlm.nih.gov/24288022/

40. Ballard O, Morrow AL. Human milk composition: nutrients and bioactive factors. Pediatr Clin North Am [Internet]. 2013 Feb [cited 2022 Jul 27];60(1):49–74. Available from: https://pubmed.ncbi.nlm.nih.gov/23178060/

41. Garlick PJ. Protein requirements of infants and children. Nestle Nutrition workshop series Paediatric programme [Internet]. 2006 [cited 2022 Jul 27];58(58). Available from: https://pubmed.ncbi.nlm.nih.gov/16902324/

42. Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, et al. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe [Internet]. 2015 Jun 1 [cited 2022 Jul 27];17(6):852. Available from: https://pubmed.ncbi.nlm.nih.gov/26308884/

43. Stewart CJ, Ajami NJ, O'Brien JL, Hutchinson DS, Smith DP, Wong MC, et al. Temporal development of the gut microbiome in early childhood from the TEDDY study. Nature [Internet]. 2018 Oct 25 [cited 2021 Nov 11];562(7728):583–8. Available from: https://pubmed.ncbi.nlm.nih.gov/30356187/

44. Guaraldi F, Salvatori G. Effect of breast and formula feeding on gut microbiota shaping in newborns. Front Cell Infect Microbiol [Internet]. 2012 [cited 2022 Jul 27];2:94. Available from: https://pubmed.ncbi.nlm.nih.gov/23087909/

45. Jost T, Lacroix C, Braegger CP, Chassard C. New insights in gut microbiota establishment in healthy breast fed neonates. PLoS One [Internet]. 2012 Aug 30 [cited 2022 Jul 27];7(8). Available from: https://pubmed.ncbi.nlm.nih.gov/22957008/

46. van den Elsen LWJ, Garssen J, Burcelin R, Verhasselt V. Shaping the Gut Microbiota by Breastfeeding: The Gateway to Allergy Prevention? Front Pediatr [Internet]. 2019 [cited 2022 Jul 27];7(FEB). Available from: https://pubmed.ncbi.nlm.nih.gov/30873394/

47. Milani C, Lugli GA, Fontana F, Mancabelli L, Alessandri G, Longhi G, et al. METAnnotatorX2: a Comprehensive Tool for Deep and Shallow Metagenomic Data Set Analyses. mSystems [Internet]. 2021 Jun 29 [cited 2022 Jan 26];6(3). Available from: https://pubmed.ncbi.nlm.nih.gov/34184911/

48. Chen Y, Ye W, Zhang Y, Xu Y. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res [Internet]. 2015 Jul 22 [cited 2022 Jan 21];43(16):7762–8. Available from: https://pubmed.ncbi.nlm.nih.gov/26250111/

49. Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, et al. The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Research [Internet]. 2018 Jan 1 [cited 2021 Sep 21];46(D1):D633–9. Available from: https://pubmed.ncbi.nlm.nih.gov/29059334/

50. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology

2019 37:8 [Internet]. 2019 Jul 24 [cited 2022 Jul 27];37(8):852–7. Available from: https://www.nature.com/articles/s41587-019-0209-9

Chapter 10

Investigation of the Ecological Link between Recurrent Microbial Human Gut Communities and Physical Activity

Chiara Tarracchini#, Federico Fontana#, Gabriele Andrea Lugli, Leonardo Mancabelli, Giulia Alessandri, Francesca Turroni, Marco Ventura*, Christian Milani*

The results of this chapter were published in Microbiology Spectrum, 2022 Mar; doi: 10.1128/spectrum.00420-22.

*These authors contributed equally.

#These authors contributed equally.

ABSTRACT

Emerging evidence has shown an association between the composition of intestinal microbial communities and host physical activity, suggesting that modifications of the gut microbiota composition may support training, performance, and post-exercise recovery of the host. Nevertheless, investigation of differences in the gut microbiota between athletes and individuals with reduced physical activity is still lacking. In this study, we performed a meta-analysis of 207 publicly available shotgun metagenomics sequencing data of fecal samples from athletes and healthy non-athletes. Accordingly, analysis of species-level fecal microbial profiles revealed three recurring compositional patterns, named HPC1 to 3, that characterize the host based on their commitment to physical activity. Interestingly, the gut microbiome of athletes showed a higher abundance of anti-inflammatory, health-promoting bacteria than that of non-athletic individuals. Moreover, the bacterial species profiled in the gut of professional athletes are short-fatty acid producers, which potentially improve energy production, and therefore sports performances. Intriguingly, microbial interaction network analyses suggested that exercise-induced microbiota adaptation involves the whole microbial community structure, resulting in a complex microbe-microbe interplay driven by positive relationships among the predicted butyrate-producing community members.

IMPORTANCE. Through metagenomic analyses, this work revealed that athletes have a gut-associated microbial community enriched in butyrate-producing species compared with non-athletes. This evidence can support the existence of a two-way association between the host's lifestyle and the gut microbiota composition, with potential intriguing athletic performance outcomes.

For Supplementary Materials see the article published in Microbiology Spectrum

INTRODUCTION

The human gut harbors a complex community of microorganisms, commonly referred to as human gut microbiota, which is well-known to play a role in nutrient uptake, vitamin synthesis, energy harvest, inflammatory modulation, and host immune response (1,–3). In turn, numerous host-dependent factors, such as genetics, age, antibiotic use, and diet, can affect the gut microbiota resulting in a highly dynamic and individual gut ecosystem (4). Recently, it has been argued that physical activity can influence gut microbiota composition, depending on the type, intensity, and exercise duration. The gut microbiota, in return, may affect the athlete's health and performance (5). Indeed, if moderate exercises (50% to about 70% of the maximum heart rate) (6) have been reported to increase the overall gut microbiota's (bio)diversity (7), prolonged endurance exercises (70% to about 85% of the maximum heart rate) (6) have been linked with an increased abundance of gut bacterial species producing short-chain fatty acids (SCFAs) (8).

In particular, members of the *Veillonella* genus, along with the metabolic pathways that this taxon utilizes for lactate conversion to propionate, have been detected with elevated abundances in athletes (9), thereby contributing to host metabolic efficiency by increasing energy availability, and thus ultimately influencing athlete performance (10). Moreover, a recent study involving professional and competitive unprofessional cyclists showed that a high training load of the cyclists corresponds to a high abundance of gut-associated *Prevotella* genus members (11). Notably, the presence of this genus has been correlated with increased metabolism of branch chain amino acids, i.e., leucine, valine, and isoleucine (11), which stimulates muscle protein synthesis and accelerates recovery (12). Furthermore, athletes generally consume higher energy diets than sedentary individuals, maintaining a high consumption of carbohydrates and proteins and a low-fat intake, with implications in gut microbiota composition (13).

In this context, our study aimed to explore the microbial communities inhabiting the gut of athletes and non-athletic individuals to highlight compositional and structural differences at the species level. For this purpose, we performed a meta-analysis employing 207 shotgun metagenomics data sets retrieved from public repositories.

RESULTS AND DISCUSSION

Meta-analysis of athletic and non-athletic individuals: data set selection and bioinformatics

Public repositories were screened for all available shotgun metagenomic data sets of the gut microbiomes of the athletes and non-athletic individuals. Specifically, we selected fecal metagenomics data from multiple sources to avoid the limitations of a single-center study. Nevertheless, combining existing data from different studies could lead to biased results due to the different strategies used to generate data sets. In particular, while the DNA extraction method has been shown to produce a little impact on the microbial structure of samples with high microbial load (14), the diverse sequencing protocols could produce different results due to differences in sequence read length and different methodologies exploited to determine the nucleotide sequences. Accordingly, to achieve high resolution of the input data and avoid the above-mentioned bias, we focused only on metagenomic data sets obtained by Illumina sequencing platform.

In detail, shotgun metagenomic sequencing data of 207 fecal samples from 107 non-athletes and 100 athletes engaged in different types of sport (cyclist, rugby players, rower, runner, and marathon athletes) were collected from six different studies (9, 11, 15,–18) and submitted to a meta-analysis aimed at elucidating the microbial species composition (Table S1). After quality filtering and removal of reads mapping against the *Homo sapiens* genome, we obtained a collection of high-quality metagenomic samples with an average of 11,700,594 reads per sample (Table S1).

As previously suggested (7), the evaluation of the alpha-diversity, expressed as the species richness, showed statistically significant differences between the gut microbiomes of non-athletic individuals and athletes, with this latter showing a

higher intestinal microbial biodiversity (average of 30 versus 34 species with relative abundance > 0.05%, *t* test *P*-value < 0.05) (Table S2). Similarly, analysis of inter-individual variability through PCoA revealed statistically significant differences in the composition of fecal microbiota between athletes and non-athletes (PERMANOVA *P*-values < 0.05) regardless of ethnic-geographic location, gender, sport type, and study cohort (PERMANOVA *P*-values > 0.05), reflecting the notion that exercise and exercise-related factors can shape the human gut microbial communities (Fig. S1a).

Taxonomic-based sample clustering and identification of the high prevalence clusters

Hierarchical clustering (HCL) analysis was performed in combination with the Silhouette method (9), employing the species-level relative abundance data to capture recurrent different taxonomic profiles from metagenomic samples. This approach led to obtaining a statistically optimal number of 10 sample clusters based on their different bacterial composition, representing the community state types (CSTs), i.e., the recurring microbial patterns observed across the investigated cohort of individuals (Fig. 1a, Fig. S1b). Among these, three were identified as the most recurrent microbial profiles, referred to as high prevalence clusters (HPCs), covering individually at least 15% of the samples and collectively 73% of the subjects included in the meta-analysis (Fig. 1a, Table S3).

Integration of the HCL analysis with the available metadata highlighted peculiar associations between HPCs and physical activity levels. In detail, while HPC1 showed a mixed composition (52% of athletes and 48% of control individuals), HPC2 encompassed 82% of non-athletes and HPC3 included 87% of athletes (Fig. 1a, Table S3). To note, after accounting for the study of origin, only 5% of the observed inter-samples variability was explained, demonstrating that geographic location and

sample processing methods do not significantly impact on the microbial composition of the subjects included in HPC3 (Fig. S1c). Consistently, the subjects included in the above-mentioned HPCs showed statistically diverse gut microbiome composition, as evidenced by the principal coordinate analysis (PCoA) based on the microbial profiling data at the species level (Fig. 1a).

Moreover, microbial biodiversity appears significantly higher in HPC3 (87% of athletes) compared with HPC1 (75% of non-athletes) (average species-richness of 36.8 versus 31.4) (Fig. 1a). As a result, at first glance, it seems that the gut microbiota of athletes is significantly diverse and more complex in terms of taxonomic composition compared to those of subjects with a more sedentary lifestyle. In particular, through the use of a polynomial linear model, which allows assessing the variability explained by each species (indicated as Adj. R-Square), we highlighted 33 taxa with a value greater than 0.15 (19), thus representing the bacterial species having the most impact in defining HPC structures (Table S3). In detail, these high-impact taxa covered from 45.86% to 68.80% of the three HPC bacterial compositions and highlighted clear connections between specific taxonomic patterns and the host's physical activity level, as discussed below (Table S3).

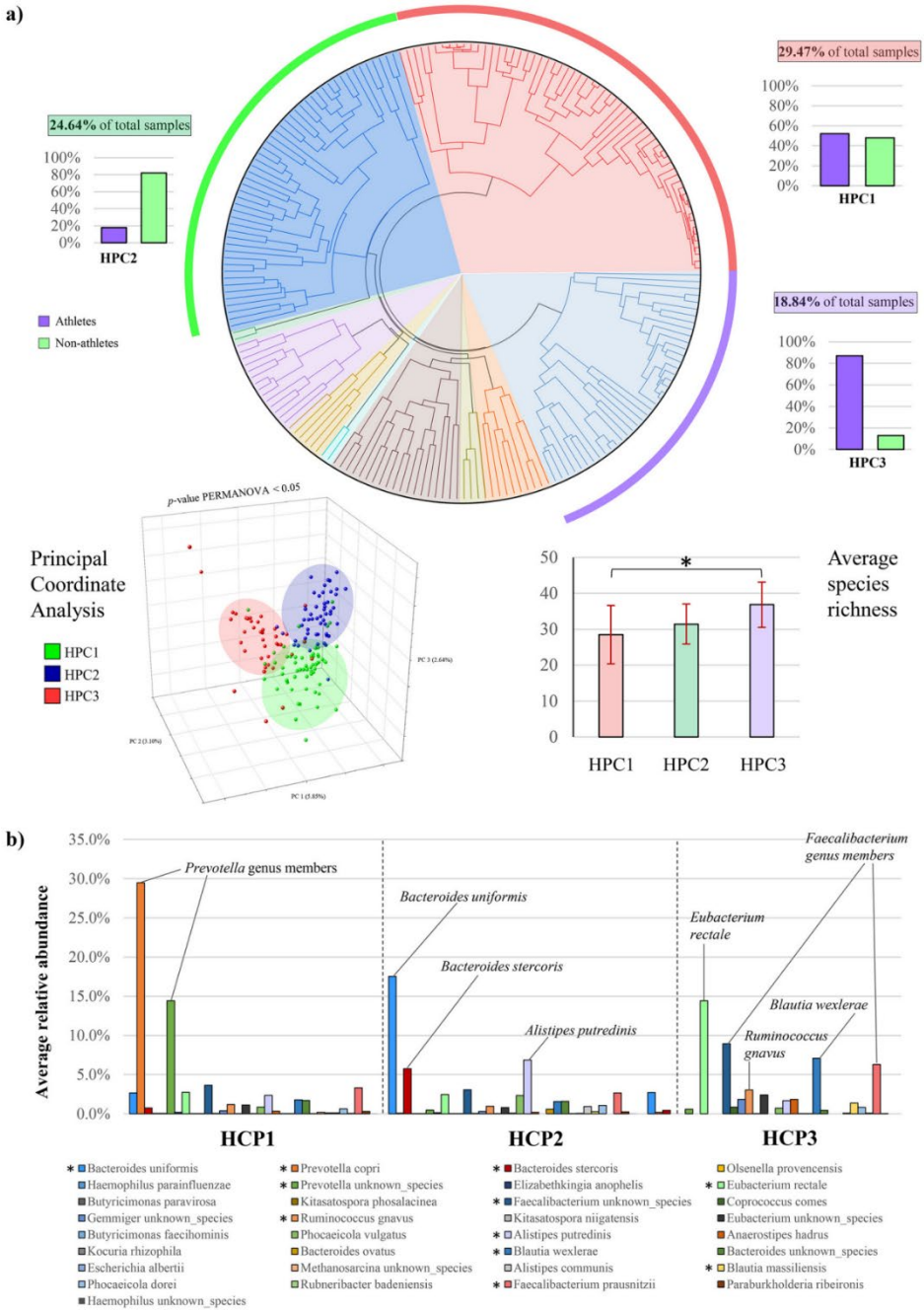


Figure 1. Cluster analysis of the 100 athletes and 107 non-athlete subjects based on gut-associated microbial community composition. Panel **(a)** shows the circular HCL-based dendrogram resulting from metagenomic sample clustering that led to the definition of the three high prevalence clusters (HPCs). The proportions of metagenomic samples from athlete and non-athletic individuals in each HPC are reported through histograms outside the circle. Below, alpha- and beta-diversity analyses involving the three HPCs are depicted through a PCoA plot and a bar chart, respectively. In panel **(b)**, the microbial taxonomic composition is visualized through a bar chart showing the average relative

abundance of each taxon at the species level. The main bacterial species showing statistically significant differences between HPCs are highlighted with asterisks on the chart legend.

Dissection of the key microbial players of the gut microbiome of athletes and non-athletic individuals

In order to catch the association between physical activity and specific taxa, we focused on the 33 microbial taxa individuated above as responsible for the main compositional differences between the three HPCs.

In particular, HPC1, composed of 52% of athletes and 48% of non-athletes, was distinct in having high relative abundances of *Prevotella* genus members (average relative abundance of 43.9%) (Fig. 1b, Table S3), which are considered a common commensal microorganism often associated with high dietary fiber intakes (20).

In contrast, HPC2, composed of 82% of samples from non-athletic individuals, was defined by the presence of *Bacteroides* members, including *Bacteroides uniformis* with an average relative abundance of 17.5% (Fig. 1b, Table S3), as expected from healthy subjects (21). Indeed, the *Bacteroides* taxon is well-known to represent a large portion of the dominant healthy human gut microbiota, previously reported to characterize one of the three renowned human enterotypes (22). Nevertheless, based on HPC2 composition, a non-athletic lifestyle was associated with increased *Alistipes putredinis* abundance (average relative abundance of 5.9%) compared with individuals with high physical activity, i.e., HPC3 (Fig. 1b, Table S3). This taxon is a member of a relatively recent genus taxonomically closely related to the *Bacteroidetes* phylum (23), whose role in the gut ecosystem is controversial (24). However, previous studies have suggested an association between *Alistipes* and

inflammation and disease, including cardiovascular disease and colorectal cancer (25, 26).

Of note, HPC3, composed for the 87% of athletes, is characterized by members of *Faecalibacterium* genus, along with *Eubacterium rectale* and *Blautia wexlerae*, with average relative abundances of 15.2%, 14.4%, and 7.1%, respectively, thus resulting significantly higher than those of non-athletic individuals (P-values < 0.05) (Fig. 1b, Table S3). Interestingly, *F. prausnitzii*, *E. rectale*, and members of the *Blautia* genus have been linked with beneficial effects in various clinical conditions, including inflammatory bowel diseases, metabolic syndromes, and colorectal cancer (27,–29). Moreover, these taxa have been reported to be responsible for butyrate production (30,–32), contributing not only to intestinal anti-inflammatory effects but also to host energy metabolism through *de novo* synthesis of glucose and lipids, which are primary sources of energy for the host organism (33, 34).

Remarkably, these findings revealed clear structural differences between the gut microbiota of the athletes and that of subjects with no physical activity, suggesting the importance of athlete gut-associated microorganisms both as supporters of the gut homeostasis as well as a source of compounds that can increase energy harvest, thus possibly improving athlete performances. However, the limited availability of precise information regarding the individual nutrition regimen did not allow further investigation of the correlation between diet and gut microbiota composition. Thus, future studies will need to collect as a wide range of metadata as possible, including dietetic regimes, that could be essential to understanding how exercise and exercise-associated factors affect the gut microbiota-host interactions in athletes.

Analysis of the interaction networks sustaining the gut microbial community of athletes and non-athletes

In order to explore the intricate interaction network of the multispecies community constituting the three HPCs, we performed a microbial co-occurrence analysis aimed at highlighting the degree of displacement (negative links) or coexistence (positive links) between species (Table S4). Correlation data were represented by a network of nodes (microbial species) linked in pairs by green edges when the relationships were positive or red edges when they were negative. Furthermore, modularity clusters (MCs) analysis allowed to detect community (sub)structures in networks, i.e., groups of taxa highly interconnected (Fig. 2, Fig. 3). Interestingly, the comparison between the network describing the gut-associated microbial community from athletes and non-athletes revealed a marked difference in the number of statistically significant interactions among taxa (positive and negative links) (Fig. 2). In particular, the microbial network of athletes showed 328 statistically significant associations, of which 62% were positive, in contrast to a total of 223 found gut microbiota members of non-athletic individuals (Table S4). Generally, compared with relatively simple networks, complex interconnected networks have a higher nutritional interaction among community members, such as cross-feeding of essential small molecules, resulting in a more stable microbial consortium with improved resilience to ecosystem disturbances (35).

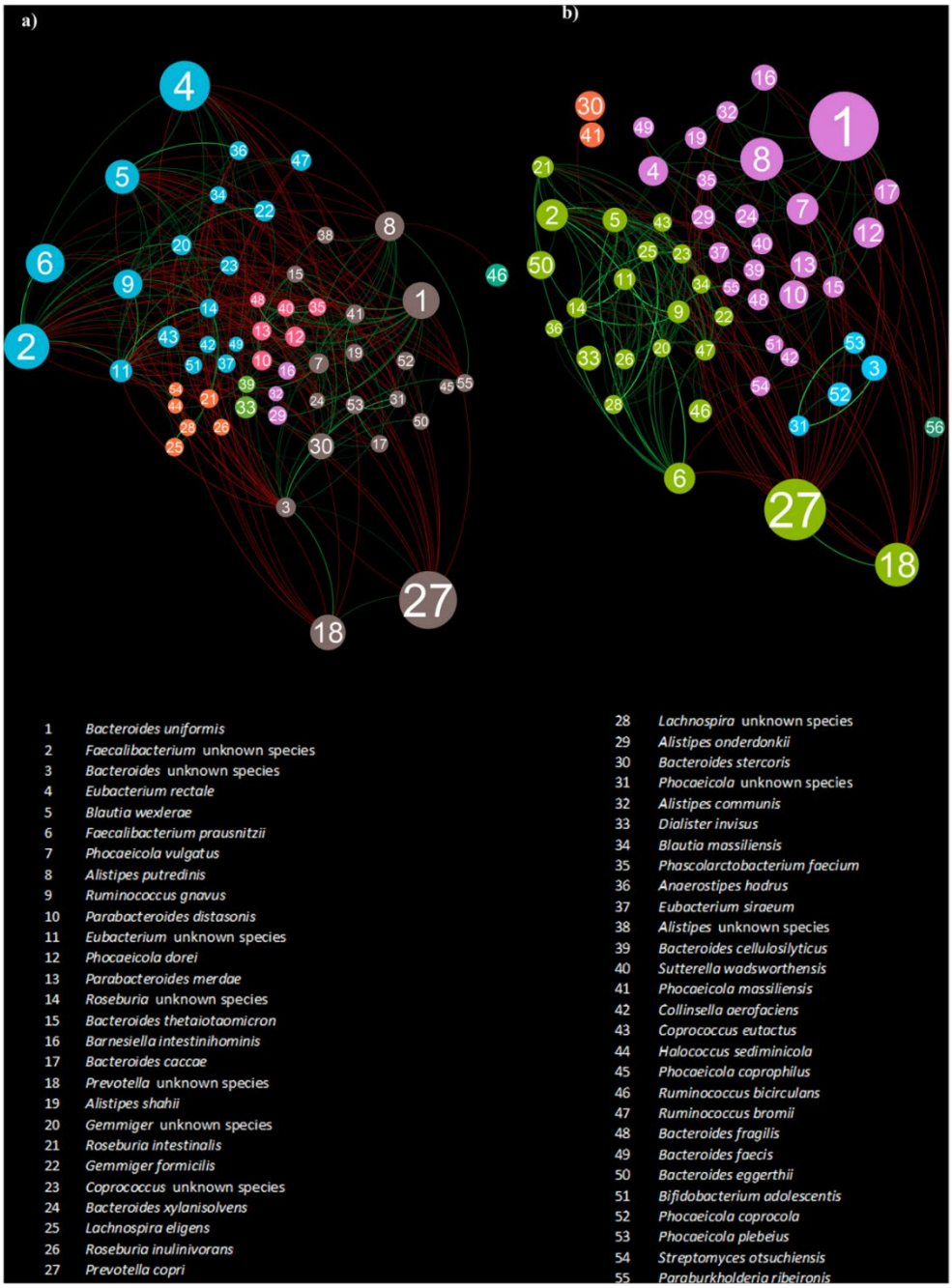


Figure 2. Interaction network supporting the structure of the gut microbial consortia in athletes and non-athletes. Panel **(a)** reports the interaction network of athlete gut microbiota, and panel **(b)** depicts the interaction network of the fecal microbial community of non-athletic individuals. In the force-driven networks, nodes represent bacterial taxa, and covariance values were used to construct the edges. Red edges correspond to negative correlations, while green edges represent positive associations. The node size is proportional to the relative average abundance of each taxon.

In addition, among the taxa with a prominent role in athlete's gut microbiota structure, we found species belonging to *Faecalibacterium*, *Eubacterium*, *Ruminococcus*, and *Blautia* genera that are thought to promote intestinal barrier integrity and prevent inflammation (36). Accordingly, these results suggested that the microbial community of athletes exhibits improved stability compared with the gut microbiome of non-athletic individuals, pointing to the importance of microbial synergism among health-promoting species in sustaining exercise-induced microbiome changes.

Co-occurrence network analyses of HPCs1 to 3

Focusing on individual HPC-derived networks, microbial correlation analysis of HPC1, which showed a mixed composition of non-athletic and athletic individuals, it is worth mentioning that members of *Prevotella* genus (node 27 and 18), such as *Prevotella copri* (node 27), tend to dominate their intestinal ecological niche. In addition, this taxon negatively correlated with other typical key members of the healthy gut-associated microbial communities, including *B. uniformis*, *Ruminococcus gnavus*, and members of *Faecalibacterium* genus (Fig. 3a, Table S4). Simultaneously, a dense and intricate network of positive associations between minority players (a proportion of 94% of the total network interactions) seems to sustain the microbial community structure of HPC1.

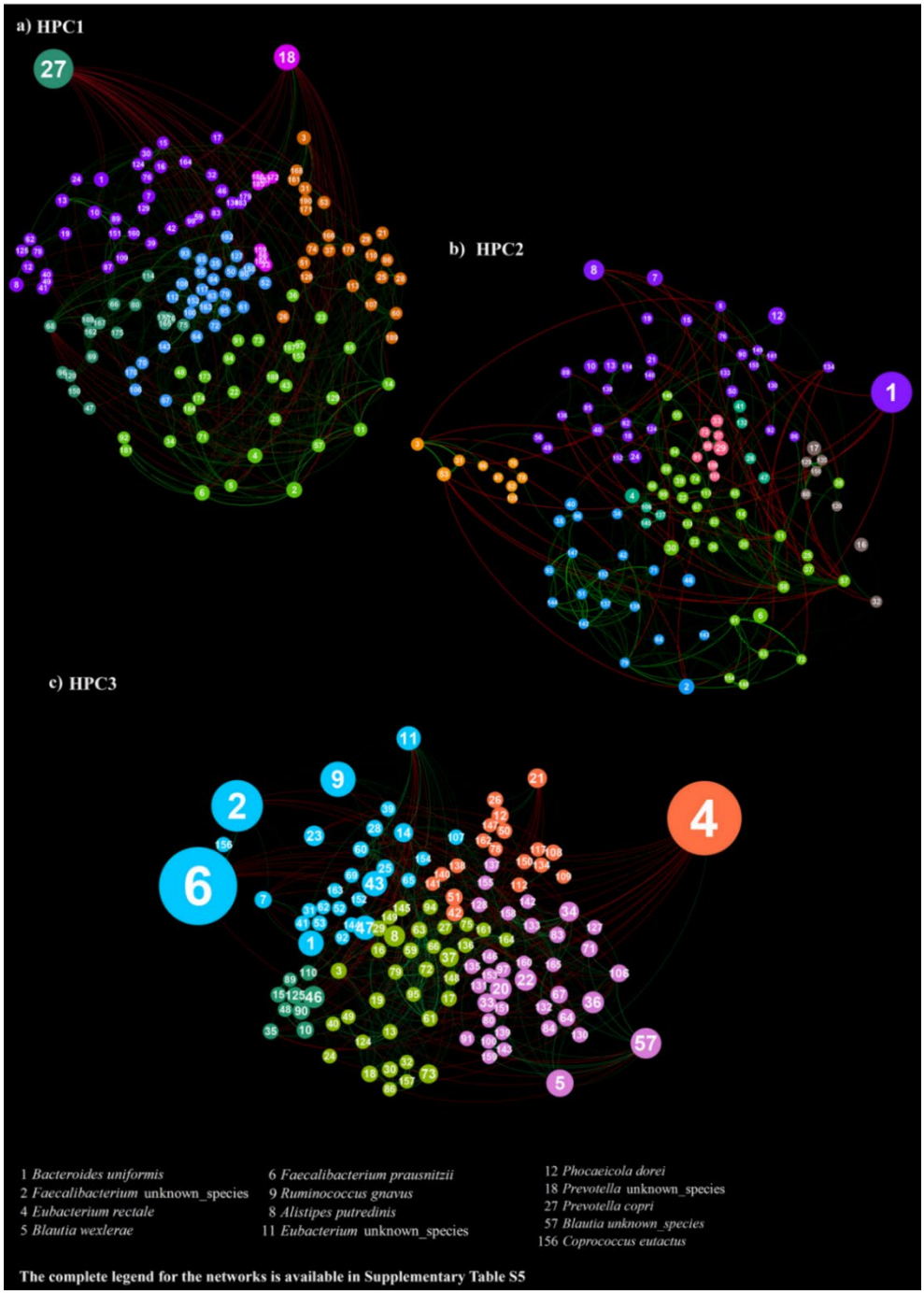


Figure 3. Co-occurrence network characterizing the three HPCs. The networks visualize the covariance relationships between the microbial taxa composing HPC1 (panel a), HPC2 (panel b), and HPC3 (panel c). HPC1 encompasses 52% of athletes and 48% of non-athletic subjects, HPC2 contains 82% of non-athletes, and HPC3 contains 87% of athletes. The complete one-to-one correspondence between node labels and microbial taxa is available in Table S5.

Conversely, the HPC2, which covers mainly non-athletic subjects, appeared to be driven by five related keystone taxa, belonging to *Bacteroides* (nodes 1 and 30), *Phocaeicola* (nodes 12 and 41), and *Alistipes* (node 8) genera (Fig. 3b, Table S4). In particular, these taxa were engaged in negative correlations mainly with potentially anti-inflammatory, butyrate-producing bacteria from the genera *Ruminoccoccus*, *Faecalibacterium*, and *Blautia* (28, 37), thus revealing a possible negative impact of a sedentary or low physical activity lifestyle on health-associated commensal bacteria. However, a small-scale subnetwork (light blue) comprising well-known commensals of the healthy human gut microbiota, such as *Bifidobacterium longum*, *Bifidobacterium adolescentis*, and *Collinsella aerofaciens* (Fig. 3b, Table S4), despite their low relative abundance in non-athletic subjects (<1%), seem to play a pivotal role in establishing positive correlations with other minor microbial players, regulating a large part of the microbial consortium characterizing healthy non-athletic individuals.

Interestingly, interaction networks describing the gut-associated microbial community of athletes, i.e., HPC3 (Fig. 3c, Table S4), showed the highest number of species that, being involved in conspicuous biotic interactions, seem to influence the whole-community dynamics of the athlete gut microbiota. Indeed, as previously mentioned, health-associated species, i.e., *Faecalibacterium prausnitzii* (node 6), *Blautia wexlerae* (node 5), and *Eubacterium rectale* (node 4), along with *Ruminococcus gnavus* (node 9), act as keystone taxa in HPC3, exerting considerable control on the entire community structure (Fig. 3c, Table S4). In addition, these taxa are involved in strong positive associations (Spearman correlation coefficient value > 0.5) with members of the *Coprococcus* and *Roseburia* genera that, being part of commensal bacteria

producing SCFAs, primarily butyrate, exert a positive influence on intestinal barrier maintenance, colonic motility, and anti-inflammatory processes (38,–40).

Besides, additional low-abundance members appear to have significant effects on the intestinal niche, reflecting the existence of a complex and solid ecosystem. As a result, removing a few species likely does not lead to a dramatic shift in the composition. Taken together, these findings support the notion that exercise can affect the gut microbiota composition, inducing qualitative and quantitative changes that may confer beneficial effects to the host and possibly to athletic performance.

CONCLUSIONS

Accumulating evidence has suggested a bidirectional association between physical activity and the composition of the microbial communities inhabiting the human intestinal environment (41). Indeed, differences in the gut microbiota composition have been observed between athletes and non-athletes, with this latter showing an increased abundance of short-chain fatty acids (SCFAs)-producing bacterial species (8, 42). In turn, the gut microbiota is thought to play a significant role in amino acid and carbohydrate host metabolism, likely indirectly influencing athlete health, training, sports performance, and post-exercise recovery (41, 43).

In this framework, a metagenomic analysis was performed by exploiting publicly available shotgun metagenomic data sets with the aim to provide insights into the gut-associated microbial community structure in athletes. In particular, a collection of 100 metagenomic samples from athletes and 107 from healthy non-athletic individuals allowed us to identify three high prevalence clusters (HPC1 to 3), i.e., recurring patterns of microbial composition. Interestingly, the gut microbiome of athletes (HPC3) showed higher biodiversity with an increased abundance of gut-associated health-promoting bacterial species compared to non-athletes.

In particular, SCFAs-producing species such as *F. prausnitzii*, *E. rectale*, *B. wexlerae*, and *R. gnavus*, were associated with athlete physical activity, revealing their possible contribution to the host health, regulating inflammation and immune system, as well as athlete's energy acquisition and sport performances. Moreover, an intricate and solid network of biotic interactions sustained by seven health-promoting key species and a range of concurrent low-abundance taxa seems to characterize the microbial community of athletes. In contrast, a less clustered and less inter-connected network was obtained from non-athletic subjects. Based on these findings, it appears that exercise induces gut microbiota changes, resulting in an increased abundance of bacteria with potential health benefits, such as SCFAs producers, cooperating in complex, interconnected microbial communities, with possible positive implications on sports performance. Future detailed functional analysis addressing the metabolic capability of the gut microbiota will aid in elucidating the connection between microbial-derived metabolites and athletic versus non-athletic lifestyles.

MATERIALS AND METDHODS

Metagenomic sample collection

With the aim to explore the differences in the gut microbiome composition between athletes and non-athletic individuals, we retrieved all the publicly available shotgun metagenomic raw data (fastq) from the National Center of Biotechnology Information (NCBI) Sequence Read Archive (SRA) database. Accordingly, to safeguard consistency and equivalence across metagenomic samples from different studies, we selected only those produced through Illumina sequencing method. As a result, we collected 207 shotgun metagenomics samples from six different studies (PRJEB15388, PRJEB28338, PRJEB32794, PRJNA472785, PRJNA305507, PRJEB2005 4), of which 100 corresponded to athlete gut microbiomes, and 107 were from healthy

non-athletes (Table S1). In addition, the respective metadata regarding health status, training type, exercise intensity level, and diet were also collected (Table S1).

Metagenomics data processing and taxonomic profiling

The fastq raw data obtained from publicly repositories were submitted to quality filtering to remove sequence reads with low-quality scores (<25). Subsequently, removal of reads mapping on the hg19 human reference genome was performed to exclude host DNA. This process allowed to achieve an average of 11,700,594 ± 9,886,096 reads per sample that were submitted to downstream analyses. The retained reads were subjected to taxonomic classification using METAnnotatorX2 bioinformatics platform (44), which performs MegaBLAST local alignment of reads (45) to the curated non-redundant sequence database of genomes retrieved from NCBI servers.

For each metagenomic sample, taxonomical biodiversity, i.e., species richness, was calculated as the number of gut-associated bacterial taxa whose sequenced reads had a relative abundance greater than 0.5%. Similarities between samples (beta-diversity) were calculated by Bray-Curtis dissimilarity based on species abundance. The range of similarities is calculated between values 0 and 1. PCoA representation of beta-diversity was performed using ORIGIN 2021 (https://www.originlab.com/2021). In the PCoA each dot represented a sample, distributed in tridimensional space according to its own bacterial composition.

The hierarchical clustering (HCL) of samples was achieved employing bacterial composition at the species level and was calculated through TMeV 4.8.1 software using Pearson correlation as a distance metric based on species-level information. The data obtained was represented by a dendrogram.

Microbial co-occurrence and network analyses

Covariance analysis involving the 332 bacterial species obtained by taxonomic profiling of the 207 metagenomic fecal samples was realized employing Kendall's tau rank covariance analysis (46). Using software Gephi (https://gephi.org/), the obtained correlation coefficients were exploited to build a force-driven network, whose nodes represent bacterial species, and edges define their relationships. The node size is related to the number of interactions of a specific microbial taxon, i.e., the node degree, while the edge color shows the type of interaction, i.e., positive (green) or negative (red).

Statistical analysis

ORIGIN 2021 (https://www.originlab.com/2021) and SPSS software (www.ibm.com/software/it/analytics/spss/) were used to compute statistical analyses. PERMANOVA analyses were performed using 1,000 permutations to assess p-values for differences among populations in PCoA analyses. Furthermore, bacterial abundance differences were tested by t-test analysis.

Data availability

Not applicable

ACKNOWLEDGMENTS

We thank GenProbio Srl for the financial support of the Laboratory of Probiogenomics. Part of this research is conducted using the High Performance Computing facility of the University of Parma.

References

1. Rowland I, Gibson G, Heinken A, Scott K, Swann J, Thiele I, Tuohy K. 2018. Gut microbiota functions: metabolism of nutrients and other food components. Eur J Nutr 57:1–24. doi: 10.1007/s00394-017-1445-8. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

2. Yoo JY, Groer M, Dutra SVO, Sarkar A, McSkimming DI. 2020. Gut microbiota and immune system interactions. Microorganisms 8:1587–1522. doi: 10.3390/microorganisms8101587. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

3. Hakansson A, Molin G. 2011. Gut microbiota and inflammation. Nutrients 3:637–682. doi: 10.3390/nu3060637. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

4. Hasan N, Yang H. 2019. Factors affecting the composition of the gut microbiota, and its modulation. PeerJ 7. doi: 10.7717/peerj.7502. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

5. Hughes RL, Holscher HD. 2021. Fueling gut microbes: a review of the interaction between diet, exercise, and the gut microbiota in athletes. Adv Nutr 12:2190–2215. doi: 10.1093/advances/nmab077. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

6. Piercy KL, Troiano RP, Ballard RM, Carlson SA, Fulton JE, Galuska DA, George SM, Olson RD. 2018. The physical activity guidelines for Americans. JAMA 320:2020–2028. doi: 10.1001/jama.2018.14854. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

7. Clarke SF, Murphy EF, O'Sullivan O, Lucey AJ, Humphreys M, Hogan A, Hayes P, O'Reilly M, Jeffery IB, Wood-Martin R, Kerins DM, Quigley E, Ross RP, O'Toole PW, Molloy MG, Falvey E, Shanahan F, Cotter PD. 2014. Exercise and associated dietary extremes impact on gut microbial diversity. Gut 63:1913–1920. doi: 10.1136/gutjnl-2013-306541. [PubMed] [CrossRef] [Google Scholar]

8. Hughes RL. 2020. A review of the role of the gut microbiome in personalized sports nutrition. Front Nutr 6. doi: 10.3389/fnut.2019.00191. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

9. Scheiman J, Luber JM, Chavkin TA, MacDonald T, Tung A, Pham L-D, Wibowo MC, Wurth RC, Punthambaker S, Tierney BT, Yang Z, Hattab MW, Avila-Pacheco J, Clish CB, Lessard S, Church GM, Kostic AD. 2019. Meta-omics analysis of elite athletes identifies a performance-enhancing microbe that functions via lactate metabolism. Nat Med 25:1104–1109. doi: 10.1038/s41591-019-0485-4. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

10. Turpin-Nolan SM, Joyner MJ, Febbraio MA. 2019. Can microbes increase exercise performance in athletes? Nat Rev Endocrinol 15:629–630. doi: 10.1038/s41574-019-0250-2. [PubMed] [CrossRef] [Google Scholar]

11. Petersen LM, Bautista EJ, Nguyen H, Hanson BM, Chen L, Lek SH, Sodergren E, Weinstock GM. 2017. Community characteristics of the gut microbiomes of competitive cyclists. Microbiome 5:98. doi: 10.1186/s40168-017-0320-4. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

12. Blomstrand E, Eliasson J, Karlsson HKR, Köhnke R. 2006. Branched-chain amino acids activate key enzymes in protein synthesis after physical exercise. The J Nutrition 136:269S–273S. doi: 10.1093/jn/136.1.269S. [PubMed] [CrossRef] [Google Scholar]

13. Spriet LL. 2019. Performance Nutrition for Athletes. Sports Medicine (Auckland, NZ) 49. https://pubmed.ncbi.nlm.nih.gov/30671901/. [PMC free article] [PubMed] [Google Scholar]

14. Sui H-y, Weil AA, Nuwagira E, Qadri F, Ryan ET, Mezzari MP, Phipatanakul W, Lai PS. 2020. Impact of DNA extraction method on variation in human and built environment microbial community and functional profiles assessed by shotgun metagenomics sequencing. Front Microbiol 11. doi: 10.3389/fmicb.2020.00953. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

15. Barton W, Penney NC, Cronin O, Garcia-Perez I, Molloy MG, Holmes E, Shanahan F, Cotter PD, O'Sullivan O. 2018. The microbiome of professional athletes differs from that of more sedentary subjects in composition and particularly at the functional metabolic level. Gut 67:625–633. https://pubmed.ncbi.nlm.nih.gov/28360096/. [PubMed] [Google Scholar]

16. O'Donovan CM, Connor B, Madigan SM, Cotter PD, O'Sullivan O. 2020. Instances of altered gut microbiomes among Irish cricketers over periods of travel in the lead up to the 2016 World Cup: a sequencing analysis. Travel Medicine and Infectious Dis 35:101553. doi: 10.1016/j.tmaid.2020.101553. [PubMed] [CrossRef] [Google Scholar]

17. O'Donovan CM, Madigan SM, Garcia-Perez I, Rankin A, O'Sullivan O, Cotter PD. 2020. Distinct microbiome composition and metabolome exists across subgroups of elite Irish athletes. J Sci Med Sport 23:63–68. doi: 10.1016/j.jsams.2019.08.290. [PubMed] [CrossRef] [Google Scholar]

18. Cronin O, Barton W, Skuse P, Penney NC, Garcia-Perez I, Murphy EF, Woods T, Nugent H, Fanning A, Melgar S, Falvey EC, Holmes E, Cotter PD, O'Sullivan O, Molloy MG, Shanahan F. 2018. A prospective metagenomic and metabolomic analysis of the impact of exercise and/or whey protein supplementation on the gut microbiome of sedentary adults. mSystems 3. doi: 10.1128/mSystems.00044-18. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

19. Kelsey CM, Prescott S, McCulloch JA, Trinchieri G, Valladares TL, Dreisbach C, Alhusen J, Grossmann T. 2021. Gut microbiota composition is associated with newborn functional brain connectivity and behavioral temperament. Brain Behav Immun 91:472–486. doi: 10.1016/j.bbi.2020.11.003. [PubMed] [CrossRef] [Google Scholar]

20. Kovatcheva-Datchary P, Nilsson A, Akrami R, Lee YS, De Vadder F, Arora T, Hallen A, Martens E, Björck I, Bäckhed F. 2015. Dietary fiber-induced improvement in glucose metabolism is associated with increased abundance of Prevotella. Cell Metab 22:971–982. doi: 10.1016/j.cmet.2015.10.001. [PubMed] [CrossRef] [Google Scholar]

21. Zafar H, Saier MH. 2021. Gut Bacteroides species in health and disease. Gut Microbes 13:1–20. https://pubmed.ncbi.nlm.nih.gov/33535896/. [PMC free article] [PubMed] [Google Scholar]

22. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto J-M, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, de Vos WM, Brunak S, Doré J, Weissenbach J, Ehrlich SD, Bork P, MetaHIT Consortium (additional members). 2011. Enterotypes of the human gut microbiome. Nature 473:174–180. doi: 10.1038/nature09944. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

23. Rautio M, Eerola E, Väisänen-Tunkelrott M-L, Molitoris D, Lawson P, Collins MD, Jousimies-Somer H. 2003. Reclassification of Bacteroides putredinis (Weinberg et al., 1937) in a new genus Alistipes gen. nov., as Alistipes putredinis comb. nov., and description of Alistipes finegoldii sp. nov., from human sources. Syst Appl Microbiol 26:182–188. doi: 10.1078/072320203322346029. [PubMed] [CrossRef] [Google Scholar]

24. Parker BJ, Wearsch PA, Veloo ACM, Rodriguez-Palacios A. 2020. The genus alistipes: gut bacteria with emerging implications to inflammation, cancer, and mental health. Front Immunol 11:906. doi: 10.3389/fimmu.2020.00906. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

25. Jie Z, Xia H, Zhong S-L, Feng Q, Li S, Liang S, Zhong H, Liu Z, Gao Y, Zhao H, Zhang D, Su Z, Fang Z, Lan Z, Li J, Xiao L, Li J, Li R, Li X, Li F, Ren H, Huang Y, Peng Y, Li G, Wen B, Dong B, Chen J-Y, Geng Q-S, Zhang Z-W, Yang H, Wang J, Wang J, Zhang X, Madsen L, Brix S, Ning G, Xu X, Liu X, Hou Y, Jia H, He K, Kristiansen K. 2017. The gut microbiome in atherosclerotic cardiovascular disease. Nat Commun 8. doi: 10.1038/s41467-017-00900-1. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

26. Moschen AR, Gerner RR, Wang J, Klepsch V, Adolph TE, Reider SJ, Hackl H, Pfister A, Schilling J, Moser PL, Kempster SL, Swidsinski A, Orth Höller D, Weiss G, Baines JF, Kaser A, Tilg H. 2016. Lipocalin 2 protects from inflammation and tumorigenesis associated with gut microbiota alterations. Cell Host Microbe 19:455–469. doi: 10.1016/j.chom.2016.03.007. [PubMed] [CrossRef] [Google Scholar]

27. Mukherjee A, Lordan C, Ross RP, Cotter PD. 2020. Gut microbes from the phylogenetically diverse genus Eubacterium and their various contributions to gut health. Gut Microbes 12:1802866. doi: 10.1080/19490976.2020.1802866. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

28. Liu X, Mao B, Gu J, Wu J, Cui S, Wang G, Zhao J, Zhang H, Chen W. 2021. Blautia-a new functional genus with potential probiotic properties? Gut Microbes 13:1–21. https://pubmed.ncbi.nlm.nih.gov/33525961/. [PMC free article] [PubMed] [Google Scholar]

29. Ferreira-Halder CV, Faria A. V d S, Andrade SS. 2017. Action and function of Faecalibacterium prausnitzii in health and disease. Best Pract Res Clin Gastroenterol 31:643–648. doi: 10.1016/j.bpg.2017.09.011. [PubMed] [CrossRef] [Google Scholar]

30. Nilsen M, Madelen Saunders C, Leena Angell I, Arntzen MØ, Lødrup Carlsen KC, Carlsen K-H, Haugen G, Heldal Hagen L, Carlsen MH, Hedlin G, Monceyron Jonassen C, Nordlund B, Maria Rehbinder E, Skjerven HO, Snipen L, Cathrine Staff A, Vettukattil R, Rudi K. 2020. Butyrate levels in the transition from an infant- to an adult-like gut microbiota correlate with bacterial networks associated with Eubacterium rectale and Ruminococcus gnavus. Genes 11:1245–1215. doi: 10.3390/genes11111245. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

31. Morrison DJ, Preston T. 2016. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. Gut Microbes 7:189–200. doi: 10.1080/19490976.2015.1134082. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

32. Vacca M, Celano G, Calabrese FM, Portincasa P, Gobbetti M, de Angelis M. 2020. The controversial role of human gut lachnospiraceae. Microorganisms [Internet] 8:573. doi: 10.3390/microorganisms8040573. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

33. den Besten G, van Eunen K, Groen AK, Venema K, Reijngoud DJ, Bakker BM. 2013. The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. J Lipid Res 54:2325–2340. doi: 10.1194/jlr.R036012. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

34. den Besten G, Lange K, Havinga R, van Dijk TH, Gerding A, van Eunen K. 2013. Gut-derived short-chain fatty acids are vividly assimilated into host carbohydrates and lipids. American J

Physiology Gastrointestinal and Liver Physiology 305. doi: 10.1152/ajpgi.00265.2013. [PubMed] [CrossRef] [Google Scholar]

35. Wagg C, Schlaeppi K, Banerjee S, Kuramae EE, van der Heijden MGA. 2019. Fungal-bacterial diversity and microbiome complexity predict ecosystem functioning. Nat Commun 10. doi: 10.1038/s41467-019-12798-y. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

36. Lordan C, Thapa D, Ross RP, Cotter PD. 2020. Potential for enriching next-generation health-promoting gut bacteria through prebiotics and other dietary components. Gut Microbes 11:1–20. doi: 10.1080/19490976.2019.1613124. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

37. Takahashi K, Nishida A, Fujimoto T, Fujii M, Shioya M, Imaeda H, Inatomi O, Bamba S, Andoh A, Sugimoto M. 2016. Reduced abundance of butyrate-producing bacteria species in the fecal microbial community in Crohn's disease. Digestion 93:59–65. doi: 10.1159/000441768. [PubMed] [CrossRef] [Google Scholar]

38. Nie K, Ma K, Luo W, Shen Z, Yang Z, Xiao M, Tong T, Yang Y, Wang X. 2021. Roseburia intestinalis: a beneficial gut organism from the discoveries in genus and species. Front Cell Infect Microbiol 11:757718. doi: 10.3389/fcimb.2021.757718. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

39. Valles-Colomer M, Falony G, Darzi Y, Tigchelaar EF, Wang J, Tito RY, Schiweck C, Kurilshikov A, Joossens M, Wijmenga C, Claes S, Van Oudenhove L, Zhernakova A, Vieira-Silva S, Raes J. 2019. The neuroactive potential of the human gut microbiota in quality of life and depression. Nat Microbiol 4:623–632. doi: 10.1038/s41564-018-0337-x. [PubMed] [CrossRef] [Google Scholar]

40. Canani RB, di Costanzo M, Leone L, Pedata M, Meli R, Calignano A. 2011. Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. World J Gastroenterol 17:1519–1528. doi: 10.3748/wjg.v17.i12.1519. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

41. Aya V, Flórez A, Perez L, Ramírez JD. 2021. Association between physical activity and changes in intestinal microbiota composition: A systematic review. PLoS One 16:e0247039. doi: 10.1371/journal.pone.0247039. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

42. Mohr AE, Jäger R, Carpenter KC, Kerksick CM, Purpura M, Townsend JR, West NP, Black K, Gleeson M, Pyne DB, Wells SD, Arent SM, Kreider RB, Campbell BI, Bannock L, Scheiman J, Wissent CJ, Pane M, Kalman DS, Pugh JN, Ortega-Santos CP, ter Haar JA, Arciero PJ, Antonio J. 2020. The athletic gut microbiota. J Int Soc Sports Nutr 17. doi: 10.1186/s12970-020-00353-w. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

43. Koh A, de Vadder F, Kovatcheva-Datchary P, Bäckhed F. 2016. From dietary fiber to host physiology: short-chain fatty acids as key bacterial metabolites. Cell 165:1332–1345. doi: 10.1016/j.cell.2016.05.041. [PubMed] [CrossRef] [Google Scholar]

44. Milani C, Lugli GA, Fontana F, Mancabelli L, Alessandri G, Longhi G, Anzalone R, Viappiani A, Turroni F, van Sinderen D, Ventura M. 2021. METAnnotatorX2: a comprehensive tool for deep and shallow metagenomic data set analyses. mSystems 6. doi: 10.1128/mSystems.00583-21. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

45. Chen Y, Ye W, Zhang Y, Xu Y. 2015. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res 43:7762–7768. doi: 10.1093/nar/gkv784. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

46. Liu X, Ning J, Cheng Y, Huang X, Li R. 2019. A flexible and robust method for assessing conditional association and conditional concordance. Stat Med 38:3656–3668. doi: 10.1002/sim.8202. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

Chapter 11

General Conclusions

Advances in understanding the ecology of the early-life human gut microbiota

The communities of microbes that populate the human gut are believed to be assembled within a developmental window comprising the first three years after birth. This period encompasses the first contact between host and microbe pioneers, representing a crucial phase for establishing the wider microbial community characterizing adulthood. Recently, extensive research efforts have been directed toward the origin of the microbial pioneers, their persistence within the gut environment, and how perinatal external and host-associated factors may interrupt the progression and maintenance of homeostatic interaction between the host and its gut microbiota.

A matter of particular interest is represented by the mechanisms underlying the maternal inheritance of specific gut commensals, which is known to be a key route for the first seeding of infant gut microbiota members. Specifically, it has been observed that members of the *Bifidobacterium* genus are the most frequently vertically transmitted bacteria at birth from mother's gut to infant, rapidly becoming the dominant group of the infant intestinal niche.

In this context, our extensive longitudinal meta-analysis showed that members of *Bifidobacterium longum* subsp. *longum* and *Bifidobacterium bifidum* strains can persist from infancy to adulthood because of the expression of microbial enzymes directed at metabolizing host-derived glycans (Chapter 3). However, such long-lasting persistence was observed to a much greater extent in the gut of women compared to males, probably because of a unique interaction between microbial host-glycan degrading genes and host's mucin structures. Indeed, the action of sex hormones could lead to a female-specific intestinal mucus configuration, which could effectively sustain the growth of mucin-metabolizing bifidobacterial strains

(Chapter 3). Overall, this represents an interesting outcome of ancient host-microbe coevolution, aimed at ensuring the persistence of selected (bifido)bacteria strains that may be, at some point, maternally transmitted to new generations as key pioneers of the infant gut.

It is worth mentioning that, if in the adult intestine bifidobacterial survival relies on mucin and diet-derived glycans utilization, in the gut of breastfed infants, the dominance of *Bifidobacterium* genus members is based on the metabolism of specific polysaccharide components of the human breast milk, referred to as Human Milk Oligosaccharides (HMOs). Specifically, *B. longum* subsp. *infantis* taxon is known to be the most efficient HMO-utilizer of infant gut-associated (bifido)bacteria. In this regard, based on our genomic exploration, the unique and extensive *B. longum* subsp. *infantis* gene repertoire for HMO utilization may have been obtained through horizontal gene transfer events from co-colonizing bacteria throughout the course of (sub)species evolution (Chapter 4).

Alongside the maternal gut, other body sites such as the birth canal also harbor microbes that can contribute to shaping the vaginal delivered infants' gut microbiota. The healthy vaginal microbiome is commonly characterized by a low degree of biodiversity and dominance of various *Lactobacillus* genus members, including *Lactobacillus crispatus*. This latter has attracted increasing interest because of its renowned association with healthy status, fertility, and favorable pregnancy outcomes, which has opened the possibility of its use in probiotic therapies before and during pregnancy. Nevertheless, our comparative genome analyses of *L. crispatus* strains revealed a noticeable intra-species genetic (micro)diversity, which appears reflected at the phenotypical level, with marked differences in growth performance and abilities to successfully dominate the female reproductive tract (Chapter 5). Such differences, possibly imputed to single nucleotide polymorphisms

in growth-related genetic traits, should be considered in formulating novel therapeutic products.

In contrast to *L. crispatus*, used as a typical biomarker of healthy cervicovaginal microbiota, *Gardnerella vaginalis* is regarded as the most common causative microorganism of bacterial vaginosis (BV), a non-inflammatory syndrome characterized by microecologic imbalance. Clinical studies have demonstrated the relationship between *G. vaginalis*-associated BV and adverse obstetric and gynecologic outcomes, including preterm labor and delivery. However, *G. vaginalis* was also isolated from women without BV symptoms. In accordance with our data gathered from genomic and phylogenomic analyses, *G. vaginalis* species appears to comprise nine different genotypes, with an uneven distribution of virulence-associated features, implying the existence of both pathogenic and commensal strains (Chapter 6). Thus, it seems plausible that *G. vaginalis* is also a component of normal vaginal microbiota, emphasizing the necessity to identify, in the context of BV diagnosis, the genotypes with the highest putative virulence capability and potential adverse health outcomes.

Another key research area is represented by what influences optimal versus altered infant gut microbiota structure, as the latter has been associated with short- and long-term adverse health outcomes, such as sepsis, immune diseases, asthma, and obesity. Many studies have identified several perinatal factors that can profoundly affect the establishment of the initial gut microbial community, including mode of delivery, feeding type, antibiotic use, and gestational age at birth. For instance, compared to healthy infants born at term, those born prematurely (<37 completed gestational weeks) showed an altered gut microbiome composition. This circumstance has been linked with an increased risk of Necrotizing EnteroColitis (NEC), a serious gastrointestinal disease involving sepsis and necrosis of intestinal portions. In

accordance with the implication of gut microbiota in NEC onset, our investigations identified C*lostridium neonatale* and *Clostridium perfringens,* as well as microbial-derived DL-lactate, as potential biofunctional markers for the early diagnosis of this disease (Chapter 7).

Generally, a common signature among neonatal gut dysbiosis-associated diseases is the reduced colonization of *Bifidobacterium* genus members, which are well-known for their beneficial and protective effects on infant health. For these reasons, diet supplementation with members of *Bifidobacterium* genus is increasingly used in attempts to beneficially modulate the altered infant gut microbiota. In this context, our Integrated Probiotic Database (IPDB) represents a useful tool for selecting beneficial (bifido)bacteria based on molecular biology and genetics (Chapter 8). Indeed, eubiosis of the gut microbiota can support the development and maintenance of host's metabolic functionalities in both early and later stages of life by producing thousands of unique bioactive small molecules in accordance with the host's physiological needs (Chapter 9). Interestingly, such microbial metabolites can accumulate in the intestine or reach organs and tissues through the blood circulatory system, eventually influencing host's physiology and contributing to its metabolism (Chapter 10).

Therefore, recognizing the importance of the critical formative years of the human gut microbiome and enhancing our understanding of its implication on host's functionality is essential for promoting not only the immediate health but also the long-lasting well-being of the individual.

References

1. Turunen, J. *et al.* Presence of distinctive microbiome in the first-pass meconium of newborn infants. *Sci Rep* **11**, (2021).
2. Stinson, L. F., Boyce, M. C., Payne, M. S. & Keelan, J. A. The Not-so-Sterile Womb: Evidence That the Human Fetus Is Exposed to Bacteria Prior to Birth. *Front Microbiol* **10**, (2019).
3. Younge, N. *et al.* Fetal exposure to the maternal microbiota in humans and mice. *JCI Insight* **4**, (2019).
4. Coscia, A., Bardanzellu, F., Caboni, E., Fanos, V. & Peroni, D. G. When a Neonate Is Born, So Is a Microbiota. *Life (Basel)* **11**, 1–28 (2021).
5. Al Alam, D. *et al.* Human Fetal Lungs Harbor a Microbiome Signature. *Am J Respir Crit Care Med* **201**, 1002–1006 (2020).
6. de Goffau, M. C. *et al.* Human placenta has no microbiome but can contain potential pathogens. *Nature* **572**, 329–334 (2019).
7. Kuperman, A. A. *et al.* Deep microbial analysis of multiple placentas shows no evidence for a placental microbiome. *BJOG* **127**, 159–169 (2020).
8. Gschwind, R. *et al.* Evidence for contamination as the origin for bacteria found in human placenta rather than a microbiota. *PLoS One* **15**, (2020).
9. Sterpu, I. *et al.* No evidence for a placental microbiome in human pregnancies at term. *Am J Obstet Gynecol* **224**, 296.e1-296.e23 (2021).
10. Leiby, J. S. *et al.* Lack of detection of a human placenta microbiome in samples from preterm and term deliveries. *Microbiome* **6**, (2018).
11. Lauder, A. P. *et al.* Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. *Microbiome* **4**, (2016).
12. Theis, K. R. *et al.* Does the human placenta delivered at term have a microbiota? Results of cultivation, quantitative real-time PCR, 16S rRNA gene sequencing, and metagenomics. *Am J Obstet Gynecol* **220**, 267.e1-267.e39 (2019).
13. Olomu, I. N. *et al.* Elimination of 'kitome' and 'splashome' contamination results in lack of detection of a unique placental microbiome. *BMC Microbiol* **20**, (2020).
14. Panzer, J. J. *et al.* Is there a placental microbiota? A critical review and re-analysis of published placental microbiota datasets. *BMC Microbiol* **23**, (2023).
15. Avershina, E. *et al.* Transition from infant- to adult-like gut microbiota. *Environ Microbiol* **18**, 2226–2236 (2016).
16. Ferretti, P. *et al.* Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host Microbe* **24**, 133-145.e5 (2018).
17. Korpela, K. *et al.* Selective maternal seeding and environment shape the human gut microbiome. *Genome Res* **28**, 561–568 (2018).

18. Maqsood, R. *et al.* Discordant transmission of bacteria and viruses from mothers to babies at birth. *Microbiome* **7**, (2019).

19. Ehrlich, A. M. *et al.* Indole-3-lactic acid associated with Bifidobacterium-dominated microbiota significantly decreases inflammation in intestinal epithelial cells. *BMC Microbiol* **20**, (2020).

20. Di Gioia, D., Aloisio, I., Mazzola, G. & Biavati, B. Bifidobacteria: their impact on gut microbiota composition and their applications as probiotics in infants. *Appl Microbiol Biotechnol* **98**, 563–577 (2014).

21. Ruiz, L., Delgado, S., Ruas-Madiedo, P., Sánchez, B. & Margolles, A. Bifidobacteria and Their Molecular Communication with the Immune System. *Front Microbiol* **8**, (2017).

22. Hidalgo-Cantabrana, C. *et al.* Bifidobacteria and Their Health-Promoting Effects. *Microbiol Spectr* **5**, (2017).

23. Yan, W., Luo, B., Zhang, X., Ni, Y. & Tian, F. Association and Occurrence of Bifidobacterial Phylotypes Between Breast Milk and Fecal Microbiomes in Mother-Infant Dyads During the First 2 Years of Life. *Front Microbiol* **12**, (2021).

24. Duranti, S. *et al.* Maternal inheritance of bifidobacterial communities and bifidophages in infants through vertical transmission. *Microbiome* **5**, (2017).

25. Asnicar, F. *et al.* Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems* **2**, (2017).

26. Wampach, L. *et al.* Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. *Nat Commun* **9**, (2018).

27. Yassour, M. *et al.* Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. *Cell Host Microbe* **24**, 146-154.e4 (2018).

28. Bäckhed, F. *et al.* Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* **17**, 690–703 (2015).

29. Pannaraj, P. S. *et al.* Association Between Breast Milk Bacterial Communities and Establishment and Development of the Infant Gut Microbiome. *JAMA Pediatr* **171**, 647–654 (2017).

30. Feehily, C. *et al.* Detailed mapping of Bifidobacterium strain transmission from mother to infant via a dual culture-based and metagenomic approach. *Nat Commun* **14**, (2023).

31. Roswall, J. *et al.* Developmental trajectory of the healthy human gut microbiota during the first 5 years of life. *Cell Host Microbe* **29**, 765-776.e3 (2021).

32. Mashima, I. & Nakazawa, F. The interaction between Streptococcus spp. and Veillonella tobetsuensis in the early stages of oral biofilm formation. *J Bacteriol* **197**, 2104–2111 (2015).

33. Ravel, J. *et al.* Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A* **108**, 4680–4687 (2011).

34. Gajer, P. *et al.* Temporal dynamics of the human vaginal microbiota. *Sci Transl Med* **4**, (2012).

35. Petrova, M. I., Lievens, E., Malik, S., Imholz, N. & Lebeer, S. Lactobacillus species as biomarkers and agents that can promote various aspects of vaginal health. *Front Physiol* **6**, (2015).

36. Gupta, S., Kakkar, V. & Bhushan, I. Crosstalk between Vaginal Microbiome and Female Health: A review. *Microb Pathog* **136**, (2019).

37. Aldunate, M. *et al.* Antimicrobial and immune modulatory effects of lactic acid and short chain fatty acids produced by vaginal microbiota associated with eubiosis and bacterial vaginosis. *Front Physiol* **6**, (2015).

38. Dols, J. A. M. *et al.* Molecular assessment of bacterial vaginosis by Lactobacillus abundance and species diversity. *BMC Infect Dis* **16**, (2016).

39. Ling, Z. *et al.* Molecular analysis of the diversity of vaginal microbiota associated with bacterial vaginosis. *BMC Genomics* **11**, (2010).

40. Han, Y., Liu, Z. & Chen, T. Role of Vaginal Microbiota Dysbiosis in Gynecological Diseases and the Potential Interventions. *Front Microbiol* **12**, (2021).

41. van Oostrum, N., De Sutter, P., Meys, J. & Verstraelen, H. Risks associated with bacterial vaginosis in infertility patients: a systematic review and meta-analysis. *Hum Reprod* **28**, 1809–1815 (2013).

42. Leitich, H. & Kiss, H. Asymptomatic bacterial vaginosis and intermediate flora as risk factors for adverse pregnancy outcome. *Best Pract Res Clin Obstet Gynaecol* **21**, 375–390 (2007).

43. Laghi, L. *et al.* Vaginal metabolic profiles during pregnancy: Changes between first and second trimester. *PLoS One* **16**, (2021).

44. Nunn, K. L. *et al.* Changes in the Vaginal Microbiome during the Pregnancy to Postpartum Transition. *Reprod Sci* **28**, 1996–2005 (2021).

45. Romero, R. *et al.* The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome* **2**, (2014).

46. Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc Natl Acad Sci U S A* **107**, 11971–11975 (2010).

47. Mortensen, M. S. *et al.* Modeling transfer of vaginal microbiota from mother to infant in early life. *Elife* **10**, 1–19 (2021).

48. Jašarević, E. *et al.* The composition of human vaginal microbiota transferred at birth affects offspring health in a mouse model. *Nat Commun* **12**, (2021).

49. Selma-Royo, M. *et al.* Perinatal environment shapes microbiota colonization and infant growth: impact on host response and intestinal function. *Microbiome* **8**, 1–19 (2020).

50. Mancabelli, L. *et al.* Multi-population cohort meta-analysis of human intestinal microbiota in early life reveals the existence of infant community state types (ICSTs). *Comput Struct Biotechnol J* **18**, 2480–2493 (2020).

51. Shao, Y. *et al.* Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117–121 (2019).

52. Madan, J. C. *et al.* Association of Cesarean Delivery and Formula Supplementation With the Intestinal Microbiome of 6-Week-Old Infants. *JAMA Pediatr* **170**, 212–219 (2016).

53. Lee, E. *et al.* Dynamics of Gut Microbiota According to the Delivery Mode in Healthy Korean Infants. *Allergy Asthma Immunol Res* **8**, 471–477 (2016).

54. Kim, G. *et al.* Delayed Establishment of Gut Microbiota in Infants Delivered by Cesarean Section. *Front Microbiol* **11**, (2020).

55. Hansen, S. *et al.* Birth by cesarean section in relation to adult offspring overweight and biomarkers of cardiometabolic risk. *Int J Obes (Lond)* **42**, 15–19 (2018).

56. Keag, O. E., Norman, J. E. & Stock, S. J. Long-term risks and benefits associated with cesarean delivery for mother, baby, and subsequent pregnancies: Systematic review and meta-analysis. *PLoS Med* **15**, (2018).

57. Mitselou, N. *et al.* Cesarean delivery, preterm birth, and risk of food allergy: Nationwide Swedish cohort study of more than 1 million children. *J Allergy Clin Immunol* **142**, 1510-1514.e2 (2018).

58. Ma, J. *et al.* Comparison of gut microbiota in exclusively breast-fed and formula-fed babies: a study of 91 term infants. *Sci Rep* **10**, (2020).

59. van den Elsen, L. W. J., Garssen, J., Burcelin, R. & Verhasselt, V. Shaping the Gut Microbiota by Breastfeeding: The Gateway to Allergy Prevention? *Front Pediatr* **7**, (2019).

60. Van Elburg, R. M., Fetter, W. P. F., Bunkers, C. M. & Heymans, H. S. A. Intestinal permeability in relation to birth weight and gestational and postnatal age. *Arch Dis Child Fetal Neonatal Ed* **88**, (2003).

61. Van Elburg, R. M. *et al.* Minimal enteral feeding, fetal blood flow pulsatility, and postnatal intestinal permeability in preterm infants with intrauterine growth retardation. *Arch Dis Child Fetal Neonatal Ed* **89**, (2004).

62. Ramasethu, J. Prevention and treatment of neonatal nosocomial infections. *Matern Health Neonatol Perinatol* **3**, (2017).

63. Rao, C. *et al.* Multi-kingdom ecological drivers of microbiota assembly in preterm infants. *Nature* **591**, 633–638 (2021).

64. Gibson, M. K. *et al.* Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. *Nat Microbiol* **1**, (2016).

65. Dibartolomeo, M. E. & Claud, E. C. The Developing Microbiome of the Preterm Infant. *Clin Ther* **38**, 733–739 (2016).

66. Gregory, K. E. *et al.* Influence of maternal breast milk ingestion on acquisition of the intestinal microbiome in preterm infants. *Microbiome* **4**, 68 (2016).

67. Zeissig, S. & Blumberg, R. S. Life at the beginning: perturbation of the microbiota by antibiotics in early life and its role in health and disease. *Nat Immunol* **15**, 307–310 (2014).

68. Gasparrini, A. J. *et al.* Persistent metagenomic signatures of early-life hospitalization and antibiotic treatment in the infant gut microbiota and resistome. *Nat Microbiol* **4**, 2285–2297 (2019).

69. Zwittink, R. D. *et al.* Association between duration of intravenous antibiotic administration and early-life microbiota development in late-preterm infants. *Eur J Clin Microbiol Infect Dis* **37**, 475–483 (2018).

70. Corvaglia, L. *et al.* Influence of Intrapartum Antibiotic Prophylaxis for Group B Streptococcus on Gut Microbiota in the First Month of Life. *J Pediatr Gastroenterol Nutr* **62**, 304–308 (2016).

71. Nogacka, A. *et al.* Impact of intrapartum antimicrobial prophylaxis upon the intestinal microbiota and the prevalence of antibiotic resistance genes in vaginally delivered full-term neonates. *Microbiome* **5**, (2017).

72. Mazzola, G. *et al.* Early Gut Microbiota Perturbations Following Intrapartum Antibiotic Prophylaxis to Prevent Group B Streptococcal Disease. *PLoS One* **11**, (2016).

73. Azad, M. B. *et al.* Impact of maternal intrapartum antibiotics, method of birth and breastfeeding on gut microbiota during the first year of life: a prospective cohort study. *BJOG* **123**, 983–993 (2016).

74. Stearns, J. C. *et al.* Intrapartum antibiotics for GBS prophylaxis alter colonization patterns in the early infant gut microbiome of low risk infants. *Sci Rep* **7**, (2017).

75. Sprockett, D., Fukami, T. & Relman, D. A. Role of priority effects in the early-life assembly of the gut microbiota. *Nat Rev Gastroenterol Hepatol* **15**, 197–205 (2018).

76. Chichlowski, M., Shah, N., Wampler, J. L., Wu, S. S. & Vanderhoof, J. A. Bifidobacterium longum subspecies infantis (B. infantis) in pediatric nutrition: Current state of knowledge. *Nutrients* **12**, (2020).

77. O'Neill, I., Schofield, Z. & Hall, L. J. Exploring the role of the microbiota member Bifidobacterium in modulating immune-linked diseases. *Emerg Top Life Sci* **1**, 333–349 (2017).

78. Turroni, F. *et al.* Bifidobacterium bifidum: A Key Member of the Early Human Gut Microbiota. *Microorganisms* **7**, (2019).

79. Turroni, F. *et al.* The infant gut microbiome as a microbial organ influencing host well-being. *Ital J Pediatr* **46**, (2020).

80. Sela, D. A. Bifidobacterial utilization of human milk oligosaccharides. *Int J Food Microbiol* **149**, 58–64 (2011).

81. Turroni, F. *et al.* Deciphering bifidobacterial-mediated metabolic interactions and their impact on gut microbiota by a multi-omics approach. *ISME J* **10**, 1656–1668 (2016).

82. Egan, M., Motherway, M. O. C., Ventura, M. & van Sinderen, D. Metabolism of sialic acid by Bifidobacterium breve UCC2003. *Appl Environ Microbiol* **80**, 4414–4426 (2014).

83. Luo, Y. *et al.* The role of mucin and oligosaccharides via cross-feeding activities by Bifidobacterium: A review. *Int J Biol Macromol* **167**, 1329–1337 (2021).

84. LeBlanc, J. G. *et al.* Bacteria as vitamin suppliers to their host: a gut microbiota perspective. *Curr Opin Biotechnol* **24**, 160–168 (2013).

85. Bordenstein, S. R. & Theis, K. R. Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes. *PLoS Biol* **13**, (2015).

86. Gilbert, S. F., Sapp, J. & Tauber, A. I. A symbiotic view of life: we have never been individuals. *Q Rev Biol* **87**, 325–341 (2012).

87. Koenig, J. E. *et al.* Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci U S A* **108 Suppl 1**, 4578–4585 (2011).

88. Fallani, M. *et al.* Determinants of the human infant intestinal microbiota after the introduction of first complementary foods in infant samples from five European centres. *Microbiology (N Y)* **157**, 1385–1392 (2011).

89. Beller, L. *et al.* Successional Stages in Infant Gut Microbiota Maturation. *mBio* **12**, (2021).

90. Wernroth, M. L. *et al.* Development of gut microbiota during the first 2 years of life. *Sci Rep* **12**, (2022).

91. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).

92. Savage, J. H. *et al.* Diet during Pregnancy and Infancy and the Infant Intestinal Microbiome. *J Pediatr* **203**, 47-54.e4 (2018).

93. Robertson, R. C., Manges, A. R., Finlay, B. B. & Prendergast, A. J. The Human Microbiome and Child Growth - First 1000 Days and Beyond. *Trends Microbiol* **27**, 131–147 (2019).

94. Romano-Keeler, J. & Weitkamp, J. H. Maternal influences on fetal microbial colonization and immune development. *Pediatr Res* **77**, 189–195 (2015).

95. Dzidic, M., Boix-Amorós, A., Selma-Royo, M., Mira, A. & Collado, M. C. Gut Microbiota and Mucosal Immunity in the Neonate. *Med Sci (Basel)* **6**, (2018).

96. Borre, Y. E. *et al.* Microbiota and neurodevelopmental windows: implications for brain disorders. *Trends Mol Med* **20**, 509–518 (2014).

97. Yang, I. *et al.* The Infant Microbiome: Implications for Infant Health and Neurocognitive Development. *Nurs Res* **65**, 76–88 (2016).

98. Jandhyala, S. M. *et al.* Role of the normal gut microbiota. *World J Gastroenterol* **21**, 8836–8847 (2015).

99. Brehin, C. *et al.* Evolution of gut microbiome and metabolome in suspected necrotizing enterocolitis: A case-control study. *J Clin Med* **9**, 1–13 (2020).

100. Zozaya, C. *et al.* Incidence, Treatment, and Outcome Trends of Necrotizing Enterocolitis in Preterm Infants: A Multicenter Cohort Study. *Front Pediatr* **8**, (2020).

101. Pammi, M. *et al.* Intestinal dysbiosis in preterm infants preceding necrotizing enterocolitis: A systematic review and meta-analysis. *Microbiome* **5**, (2017).

102. Wandro, S. *et al.* The Microbiome and Metabolome of Preterm Infant Stool Are Personalized and Not Driven by Health Outcomes, Including Necrotizing Enterocolitis and Late-Onset Sepsis. *mSphere* **3**, (2018).

103. Niemarkt, H. J. *et al.* Necrotizing Enterocolitis, Gut Microbiota, and Brain Development: Role of the Brain-Gut Axis. *Neonatology* **115**, 423–431 (2019).

104. Ficara, M. *et al.* Changes of intestinal microbiota in early life. *J Matern Fetal Neonatal Med* **33**, 1036–1043 (2020).

105. Shane, A. L., Sánchez, P. J. & Stoll, B. J. Neonatal sepsis. *The Lancet* **390**, 1770–1780 (2017).

106. Camacho-Gonzalez, A., Spearman, P. W. & Stoll, B. J. Neonatal infectious diseases: evaluation of neonatal sepsis. *Pediatr Clin North Am* **60**, 367–389 (2013).

107. Cortese, F. *et al.* Early and Late Infections in Newborns: Where Do We Stand? A Review. *Pediatr Neonatol* **57**, 265–273 (2016).

108. Carl, M. A. *et al.* Sepsis from the gut: the enteric habitat of bacteria that cause late-onset neonatal bloodstream infections. *Clin Infect Dis* **58**, 1211–1218 (2014).

109. El Manouni El Hassani, S. *et al.* Profound Pathogen-Specific Alterations in Intestinal Microbiota Composition Precede Late-Onset Sepsis in Preterm Infants: A Longitudinal, Multicenter, Case-Control Study. *Clin Infect Dis* **73**, E224–E232 (2021).

110. Mai, V. *et al.* Distortions in development of intestinal microbiota associated with late onset sepsis in preterm infants. *PLoS One* **8**, (2013).

111. Lee, C. C. *et al.* Gut Dysbiosis, Bacterial Colonization and Translocation, and Neonatal Sepsis in Very-Low-Birth-Weight Preterm Infants. *Front Microbiol* **12**, (2021).

112. Liu, J. *et al.* Patterned progression of gut microbiota associated with necrotizing enterocolitis and late onset sepsis in preterm infants: a prospective study in a Chinese neonatal intensive care unit. *PeerJ* **7**, (2019).

113. Stewart, C. J. *et al.* Longitudinal development of the gut microbiome and metabolome in preterm neonates with late onset sepsis and healthy controls. *Microbiome* **5**, 75 (2017).

114. Zheng, D., Liwinski, T. & Elinav, E. Interaction between microbiota and immunity in health and disease. *Cell Research 2020 30:6* **30**, 492–506 (2020).

115. Johnson, C. C. & Ownby, D. R. The infant gut bacterial microbiota and risk of pediatric asthma and allergic diseases. *Transl Res* **179**, 60–70 (2017).

116. Wang, S. *et al.* A good start in life is important-perinatal factors dictate early microbiota development and longer term maturation. *FEMS Microbiol Rev* **44**, 763–781 (2020).

117. Johnson, C. C. & Ownby, D. R. Allergies and Asthma: Do Atopic Disorders Result from Inadequate Immune Homeostasis arising from Infant Gut Dysbiosis? *Expert Rev Clin Immunol* **12**, 379–388 (2016).

118. Fujimura, K. E. *et al.* Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. *Nat Med* **22**, 1187–1191 (2016).

119. Ta, L. D. H. *et al.* A compromised developmental trajectory of the infant gut microbiome and metabolome in atopic eczema. *Gut Microbes* **12**, 1–21 (2020).

120. Stiemsma, L. T. *et al.* Shifts in Lachnospira and Clostridium sp. in the 3-month stool microbiome are associated with preschool age asthma. *Clin Sci (Lond)* **130**, 2199–2207 (2016).

121. Stanislawski, M. A. *et al.* Gut Microbiota in the First 2 Years of Life and the Association with Body Mass Index at Age 12 in a Norwegian Birth Cohort. *mBio* **9**, (2018).

122. Vael, C., Verhulst, S. L., Nelen, V., Goossens, H. & Desager, K. N. Intestinal microflora and body mass index during the first three years of life: an observational study. *Gut Pathog* **3**, (2011).

123. Jian, C. *et al.* Early-life gut microbiota and its connection to metabolic health in children: Perspective on ecological drivers and need for quantitative approach. *EBioMedicine* **69**, (2021).

124. Chambers, E. S., Preston, T., Frost, G. & Morrison, D. J. Role of Gut Microbiota-Generated Short-Chain Fatty Acids in Metabolic and Cardiovascular Health. *Curr Nutr Rep* **7**, 198–206 (2018).

125. Alfaleh, K. & Anabrees, J. Probiotics for prevention of necrotizing enterocolitis in preterm infants. *Cochrane Database Syst Rev* **2014**, (2014).

126. Sawh, S. C., Deshpande, S., Jansen, S., Reynaert, C. J. & Jones, P. M. Prevention of necrotizing enterocolitis with probiotics: a systematic review and meta-analysis. *PeerJ* **4**, (2016).

127. Zhu, X. L. *et al.* Bifidobacterium may benefit the prevention of necrotizing enterocolitis in preterm infants: A systematic review and meta-analysis. *Int J Surg* **61**, 17–25 (2019).

128. Jacobs, S. E. *et al.* Probiotic effects on late-onset sepsis in very preterm infants: a randomized controlled trial. *Pediatrics* **132**, 1055–1062 (2013).

129. Bernaola Aponte, G., Bada Mancilla, C. A., Carreazo, N. Y. & Rojas Galarza, R. A. Probiotics for treating persistent diarrhoea in children. *Cochrane Database Syst Rev* **2013**, (2013).

130. Korpela, K. *et al.* Probiotic supplementation restores normal microbiota composition and function in antibiotic-treated and in caesarean-born infants. *Microbiome* **6**, (2018).

131. Dryl, R. & Szajewska, H. Probiotics for management of infantile colic: a systematic review of randomized controlled trials. *Arch Med Sci* **14**, 1137–1143 (2018).

132. Karkhaneh, M., Fraser, L., Jou, H. & Vohra, S. Effectiveness of probiotics in infantile colic: A rapid review. *Paediatr Child Health* **25**, 149–159 (2020).

133. Alcon-Giner, C. *et al.* Microbiota Supplementation with Bifidobacterium and Lactobacillus Modifies the Preterm Infant Gut Microbiota and Metabolome: An Observational Study. *Cell Rep Med* **1**, (2020).

134. Reyman, M. *et al.* Impact of delivery mode-associated gut microbiota dynamics on health in the first year of life. *Nature Communications 2019 10:1* **10**, 1–12 (2019).

135. Dominguez-Bello, M. G. *et al.* Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. *Nat Med* **22**, 250–253 (2016).

136. Song, S. J. *et al.* Naturalization of the microbiota developmental trajectory of Cesarean-born neonates after vaginal seeding. *Med (N Y)* **2**, 951-964.e5 (2021).

137. Korpela, K. *et al.* Maternal Fecal Microbiota Transplantation in Cesarean-Born Infants Rapidly Restores Normal Gut Microbial Development: A Proof-of-Concept Study. *Cell* **183**, 324-334.e5 (2020).

138. Wilson, B. C. *et al.* Oral administration of maternal vaginal microbes at birth to restore gut microbiome development in infants born by caesarean section: A pilot randomised placebo-controlled trial. *EBioMedicine* **69**, (2021).

139. Braegger, C. *et al.* Supplementation of infant formula with probiotics and/or prebiotics: a systematic review and comment by the ESPGHAN committee on nutrition. *J Pediatr Gastroenterol Nutr* **52**, 238–250 (2011).

140. Gibson, G. R. *et al.* Expert consensus document: The International Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on the definition and scope of prebiotics. *Nat Rev Gastroenterol Hepatol* **14**, 491–502 (2017).

141. Bych, K. *et al.* Production of HMOs using microbial hosts - from cell engineering to large scale production. *Curr Opin Biotechnol* **56**, 130–137 (2019).

142. Rad, A. H., Maleki, L. A., Kafil, H. S., Zavoshti, H. F. & Abbasi, A. Postbiotics as novel health-promoting ingredients in functional foods. *Health Promot Perspect* **10**, 3–4 (2020).

143. Salminen, S., Stahl, B., Vinderola, G. & Szajewska, H. Infant Formula Supplemented with Biotics: Current Knowledge and Future Perspectives. *Nutrients* **12**, 1–20 (2020).

144. Fan, Y. & Pedersen, O. Gut microbiota in human metabolic health and disease. *Nature Reviews Microbiology 2020 19:1* **19**, 55–71 (2020).

145. Sadler, R. *et al.* Short-Chain Fatty Acids Improve Poststroke Recovery via Immunological Mechanisms. *J Neurosci* **40**, 1162–1173 (2020).

146. Fukuda, S. *et al.* Bifidobacteria can protect from enteropathogenic infection through production of acetate. *Nature* **469**, 543–549 (2011).

147. Zelante, T. *et al.* Tryptophan catabolites from microbiota engage aryl hydrocarbon receptor and balance mucosal reactivity via interleukin-22. *Immunity* **39**, 372–385 (2013).

148. Guo, X. *et al.* Innate Lymphoid Cells Control Early Colonization Resistance against Intestinal Pathogens through ID2-Dependent Regulation of the Microbiota. *Immunity* **42**, 731–743 (2015).

149. Sillner, N. *et al.* Longitudinal Profiles of Dietary and Microbial Metabolites in Formula- and Breastfed Infants. *Front Mol Biosci* **8**, (2021).

150. Brink, L. R. *et al.* Neonatal diet alters fecal microbiota and metabolome profiles at different ages in infants fed breast milk or formula. *Am J Clin Nutr* **111**, 1190–1202 (2020).

151. Laursen, M. F. *et al.* Bifidobacterium species associated with breastfeeding produce aromatic lactic acids in the infant gut. *Nat Microbiol* **6**, 1367–1382 (2021).

152. Tsukuda, N. *et al.* Key bacterial taxa and metabolic pathways affecting gut short-chain fatty acid profiles in early life. *ISME J* **15**, 2574–2590 (2021).

153. Koh, A., De Vadder, F., Kovatcheva-Datchary, P. & Bäckhed, F. From Dietary Fiber to Host Physiology: Short-Chain Fatty Acids as Key Bacterial Metabolites. *Cell* **165**, 1332–1345 (2016).

154. Bridgman, S. L. *et al.* Childhood body mass index and associations with infant gut metabolites and secretory IgA: findings from a prospective cohort study. *Int J Obes (Lond)* **46**, 1712–1719 (2022).

155. Cryan, J. F. *et al.* The Microbiota-Gut-Brain Axis. *Physiol Rev* **99**, 1877–2013 (2019).

156. Clarke, G. *et al.* The microbiome-gut-brain axis during early life regulates the hippocampal serotonergic system in a sex-dependent manner. *Mol Psychiatry* **18**, 666–673 (2013).

157. Sharon, G. *et al.* Human Gut Microbiota from Autism Spectrum Disorder Promote Behavioral Symptoms in Mice. *Cell* **177**, 1600-1618.e17 (2019).

158. Jameson, K. G. & Hsiao, E. Y. Linking the Gut Microbiota to a Brain Neurotransmitter. *Trends Neurosci* **41**, 413–414 (2018).

159. Lu, J. & Claud, E. C. Connection between gut microbiome and brain development in preterm infants. *Dev Psychobiol* **61**, 739–751 (2019).

160. Ming, X. *et al.* A Gut Feeling: A Hypothesis of the Role of the Microbiome in Attention-Deficit/Hyperactivity Disorders. *Child Neurol Open* **5**, 2329048X18786799 (2018).

161. Tabouy, L. *et al.* Dysbiosis of microbiome and probiotic treatment in a genetic model of autism spectrum disorders. *Brain Behav Immun* **73**, 310–319 (2018).

162. Yu, M. *et al.* Aryl Hydrocarbon Receptor Activation Modulates Intestinal Epithelial Barrier Function by Maintaining Tight Junction Integrity. *Int J Biol Sci* **14**, 69–77 (2018).

163. Lee, J. H. & Lee, J. Indole as an intercellular signal in microbial communities. *FEMS Microbiol Rev* **34**, 426–444 (2010).

164. Kim, J. & Park, W. Indole: a signaling molecule or a mere metabolic byproduct that alters bacterial physiology at a high concentration? *J Microbiol* **53**, 421–428 (2015).

165. Ticho, A. L., Malhotra, P., Dudeja, P. K., Gill, R. K. & Alrefai, W. A. Bile acid receptors and gastrointestinal functions. *Liver Res* **3**, 31–39 (2019).

166. Pols, T. W. H., Noriega, L. G., Nomura, M., Auwerx, J. & Schoonjans, K. The bile acid membrane receptor TGR5: a valuable metabolic target. *Dig Dis* **29**, 37–44 (2011).

167. Tanaka, M. *et al.* The association between gut microbiota development and maturation of intestinal bile acid metabolism in the first 3 y of healthy Japanese infants. *Gut Microbes* **11**, 205–216 (2020).

168. van Best, N. *et al.* Bile acids drive the newborn's gut microbiota maturation. *Nat Commun* **11**, (2020).

169. Larabi, A. B., Masson, H. L. P. & Bäumler, A. J. Bile acids as modulators of gut microbiota composition and function. *Gut Microbes* **15**, (2023).

170. Shu, S. A. *et al.* Microbiota and Food Allergy. *Clin Rev Allergy Immunol* **57**, 83–97 (2019).

171. Rooks, M. G. & Garrett, W. S. Gut microbiota, metabolites and host immunity. *Nat Rev Immunol* **16**, 341–352 (2016).

172. Knoop, K. A. *et al.* Microbial antigen encounter during a preweaning interval is critical for tolerance to gut bacteria. *Sci Immunol* **2**, (2017).

173. Tanaka, M. & Nakayama, J. Development of the gut microbiota in infancy and its impact on health in later life. *Allergol Int* **66**, 515–522 (2017).

174. Negi, S., Das, D. K., Pahari, S., Nadeem, S. & Agrewala, J. N. Potential Role of Gut Microbiota in Induction and Regulation of Innate Immune Memory. *Front Immunol* **10**, 2441 (2019).

175. Burgueño, J. F. & Abreu, M. T. Epithelial Toll-like receptors and their role in gut homeostasis and disease. *Nature Reviews Gastroenterology & Hepatology 2020 17:5* **17**, 263–278 (2020).

176. Mallard, C. Innate Immune Regulation by Toll-Like Receptors in the Brain. *ISRN Neurol* **2012**, 1–19 (2012).

177. Valentini, M. *et al.* Immunomodulation by Gut Microbiota: Role of Toll-Like Receptor Expressed by T Cells. *J Immunol Res* **2014**, (2014).

178. Stagg, A. J. Intestinal Dendritic Cells in Health and Gut Inflammation. *Front Immunol* **9**, (2018).

179. Rutella, S. & Locatelli, F. Intestinal dendritic cells in the pathogenesis of inflammatory bowel disease. *World J Gastroenterol* **17**, 3761–3775 (2011).

180. Ramakrishna, C. *et al.* Bacteroides fragilis polysaccharide A induces IL-10 secreting B and T cells that prevent viral encephalitis. *Nature Communications 2019 10:1* **10**, 1–13 (2019).

181. Faith, J. J., Ahern, P. P., Ridaura, V. K., Cheng, J. & Gordon, J. I. Identifying gut microbe-host phenotype relationships using combinatorial communities in gnotobiotic mice. *Sci Transl Med* **6**, (2014).

182. Geuking, M. B. *et al.* Intestinal bacterial colonization induces mutualistic regulatory T cell responses. *Immunity* **34**, 794–806 (2011).

183. Eberl, G. The microbiota, a necessary element of immunity. *C R Biol* **341**, 281–283 (2018).

184. Al Nabhani, Z. *et al.* A Weaning Reaction to Microbiota Is Required for Resistance to Immunopathologies in the Adult. *Immunity* **50**, 1276-1288.e5 (2019).

185. Al Nabhani, Z. & Eberl, G. Imprinting of the immune system by the microbiota early in life. *Mucosal Immunology 2020 13:2* **13**, 183–189 (2020).

186. Smith, P. M. *et al.* The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis. *Science* **341**, 569–573 (2013).

187. Arpaia, N. *et al.* Metabolites produced by commensal bacteria promote peripheral regulatory T-cell generation. *Nature* **504**, 451–455 (2013).

188. Furusawa, Y. *et al.* Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature* **504**, 446–450 (2013).

189. Nguyen, Q. N., Himes, J. E., Martinez, D. R. & Permar, S. R. The Impact of the Gut Microbiota on Humoral Immunity to Pathogens and Vaccination in Early Infancy. *PLoS Pathog* **12**, (2016).

190. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).

191. Maier, L. *et al.* Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555**, 623–628 (2018).

192. Vich Vila, A. *et al.* Impact of commonly used drugs on the composition and metabolic function of the gut microbiota. *Nat Commun* **11**, (2020).

193. Imhann, F. *et al.* Proton pump inhibitors affect the gut microbiome. *Gut* **65**, 740–748 (2016).

194. Maier, L. *et al.* Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555**, 623–628 (2018).

195. Takagi, T. *et al.* The influence of long-term use of proton pump inhibitors on the gut microbiota: an age-sex-matched case-control study. *J Clin Biochem Nutr* **62**, 100–105 (2018).

196. Karl, P. J. *et al.* Effects of Psychological, Environmental and Physical Stressors on the Gut Microbiota. *Front Microbiol* **9**, (2018).

197. Coley, E. J. L. *et al.* Early life adversity predicts brain-gut alterations associated with increased stress and mood. *Neurobiol Stress* **15**, (2021).

198. Mackner, L. M. *et al.* Fecal microbiota and metabolites are distinct in a pilot study of pediatric Crohn's disease patients with higher levels of perceived stress. *Psychoneuroendocrinology* **111**, (2020).

199. Michels, N. *et al.* Gut microbiome patterns depending on children's psychosocial stress: Reports versus biomarkers. *Brain Behav Immun* **80**, 751–762 (2019).

200. Kaur, H., Bose, C. & Mande, S. S. Tryptophan Metabolism by Gut Microbiome and Gut-Brain-Axis: An in silico Analysis. *Front Neurosci* **13**, (2019).

201. Godoy, L. D., Rossignoli, M. T., Delfino-Pereira, P., Garcia-Cairasco, N. & Umeoka, E. H. de L. A Comprehensive Overview on Stress Neurobiology: Basic Concepts and Clinical Implications. *Front Behav Neurosci* **12**, (2018).

202. Cronin, O., Molloy, M. G. & Shanahan, F. Exercise, fitness, and the gut. *Curr Opin Gastroenterol* **32**, 67–73 (2016).

203. Bressa, C. *et al.* Differences in gut microbiota profile between women with active lifestyle and sedentary women. *PLoS One* **12**, (2017).

204. Quiroga, R. *et al.* Exercise training modulates the gut microbiota profile and impairs inflammatory signaling pathways in obese children. *Exp Mol Med* **52**, 1048–1061 (2020).

205. O'Toole, P. W. & Shiels, P. G. The role of the microbiota in sedentary lifestyle disorders and ageing: lessons from the animal kingdom. *J Intern Med* **287**, 271–282 (2020).

206. Mailing, L. J., Allen, J. M., Buford, T. W., Fields, C. J. & Woods, J. A. Exercise and the Gut Microbiome: A Review of the Evidence, Potential Mechanisms, and Implications for Human Health. *Exerc Sport Sci Rev* **47**, 75–85 (2019).

207. Dalton, A., Mermier, C. & Zuhl, M. Exercise influence on the microbiome-gut-brain axis. *Gut Microbes* **10**, 555–568 (2019).

208. Cataldi, S. *et al.* The Relationship between Physical Activity, Physical Exercise, and Human Gut Microbiota in Healthy and Unhealthy Subjects: A Systematic Review. *Biology (Basel)* **11**, (2022).

209. Cândido, F. G. *et al.* Impact of dietary fat on gut microbiota and low-grade systemic inflammation: mechanisms and clinical implications on obesity. *Int J Food Sci Nutr* **69**, 125–143 (2018).

210. Cândido, T. L. N., Bressan, J. & Alfenas, R. de C. G. Dysbiosis and metabolic endotoxemia induced by high-fat diet. *Nutr Hosp* **35**, 1432–1440 (2018).

211. Martinez, K. B., Leone, V. & Chang, E. B. Western diets, gut dysbiosis, and metabolic diseases: Are they linked? *Gut Microbes* **8**, 130–142 (2017).

212. Agus, A. *et al.* Western diet induces a shift in microbiota composition enhancing susceptibility to Adherent-Invasive E. coli infection and intestinal inflammation. *Sci Rep* **6**, (2016).

213. Las Heras, V., Melgar, S., MacSharry, J. & Gahan, C. G. M. The Influence of the Western Diet on Microbiota and Gastrointestinal Immunity. *Annu Rev Food Sci Technol* **13**, 489–512 (2022).

214. Graf, D. *et al.* Contribution of diet to the composition of the human gut microbiota. *Microb Ecol Health Dis* **26**, (2015).

215. Merra, G. *et al.* Influence of Mediterranean Diet on Human Gut Microbiota. *Nutrients* **13**, 1–12 (2020).

216. Rampelli, S. *et al.* Microbiota and lifestyle interactions through the lifespan. *Trends Food Sci Technol* **57**, 265–272 (2016).

217. Wu, G. D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105–108 (2011).

218. Claesson, M. J. *et al.* Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc Natl Acad Sci U S A* **108 Suppl 1**, 4586–4591 (2011).

219. Claesson, M. J. *et al.* Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488**, 178–184 (2012).

220. Odamaki, T. *et al.* Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. *BMC Microbiol* **16**, (2016).

221. Biagi, E. *et al.* Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PLoS One* **5**, (2010).

222. Vernaya, M., McAdam, J. & Hampton, M. D. C. Effectiveness of probiotics in reducing the incidence of Clostridium difficile-associated diarrhea in elderly patients: a systematic review. *JBI Database System Rev Implement Rep* **15**, 140–164 (2017).

223. Krüger, J. F., Hillesheim, E., Pereira, A. C. S. N., Camargo, C. Q. & Rabito, E. I. Probiotics for dementia: a systematic review and meta-analysis of randomized controlled trials. *Nutr Rev* **79**, 160–170 (2021).

224. Martínez-Martínez, M. I., Calabuig-Tolsá, R. & Cauli, O. The effect of probiotics as a treatment for constipation in elderly people: A systematic review. *Arch Gerontol Geriatr* **71**, 142–149 (2017).

225. Miller, L. E., Lehtoranta, L. & Lehtinen, M. J. Short-term probiotic supplementation enhances cellular immune function in healthy elderly: systematic review and meta-analysis of controlled studies. *Nutr Res* **64**, 1–8 (2019).

226. Rizzoli, R. & Biver, E. Are Probiotics the New Calcium and Vitamin D for Bone Health? *Curr Osteoporos Rep* **18**, 273–284 (2020).

227. Xie, C., Li, J., Wang, K., Li, Q. & Chen, D. Probiotics for the prevention of antibiotic-associated diarrhoea in older patients: a systematic review. *Travel Med Infect Dis* **13**, 128–134 (2015).

Pubblications

(articles published during my PhD training)

1. Tarracchini C., Argentini C., Alessandri G., Lugli G.A., Mancabelli L., Fontana F., Anzalone R., Viappiani A., Turroni F., Ventura M., Milani C. **"The core genome evolution of Lactobacillus crispatus as a driving force for niche competition in the human vaginal tract"**. Microb Biotechnol. 2023 Sep. (IF: 5.7).

2. Mancabelli L., Taurino G., Ticinesi A., Ciociola T., Vacondio F., Milani C., Fontana F., Lugli G.A., Tarracchini C., Alessandri G., Viappiani A., Bianchi M., Nouvenne A., Chetta A.A., Turroni F., Meschi T., Mor M., Bussolati O., Ventura M. **"Disentangling the interactions between nasopharyngeal and gut microbiome and their involvement in the modulation of COVID-19 infection".** Microbiol Spectr. 2023 Sep. (IF: 3.7).

3. Tarracchini C., Alessandri G., Fontana F., Rizzo S.M., Lugli G.A., Bianchi M.G., Mancabelli L., Longhi G., Argentini C., Vergna L.M., Anzalone R., Viappiani A., Turroni F., Taurino G., Chiu M., Arboleya S., Gueimonde M., Bussolati O., van Sinderen D., Milani C., Ventura M. **"Genetic strategies for sex-biased persistence of gut microbes across human life"**. Nat Commun. 2023 Jul 14. (IF: 17.7).

4. Rizzo S.M., Alessandri G., Lugli G.A., Fontana F., Tarracchini C., Mancabelli L., Viappiani A., Bianchi M.G., Bussolati O., van Sinderen D., Ventura M., Turroni F. **"Exploring Molecular Interactions between Human Milk Hormone Insulin and Bifidobacteria"**. Microbiol Spectr. 2023 May. (IF: 9.0).

5. Mancabelli L., Milani C., Fontana F., Liotto N., Tabasso C., Perrone M., Lugli G.A., Tarracchini C., Alessandri G., Viappiani A., Bernasconi S., Roggero P., Mosca F., Turroni F., Ventura M. **"A pilot study to disentangle the infant gut microbiota composition and identification of bacteria correlates with high fat mass"**. Microbiome Res Rep. 2023 Jun.

6. Lugli G.A., Mancabelli L., Milani C., Fontana F., Tarracchini C., Alessandri G., van Sinderen D, Turroni F., Ventura M. **"Comprehensive insights from**

composition to functional microbe-based biodiversity of the infant human gut microbiota". NPJ Biofilms Microbiomes. 2023 May (IF: 8.5).

7. Lugli G.A., Fontana F., Tarracchini C., Milani C., Mancabelli L., Turroni F., Ventura M. **"MEGAnnotator2: a pipeline for the assembly and annotation of microbial genomes"**. Microbiome Res Rep. 2023 Apr.

8. Argentini C., Tarracchini C., Alessandri G., Longhi G, Milani C., van Sinderen D., Ventura M., Turroni F. **"Contribution of the capsular polysaccharide layer to antibiotic resistance in bifidobacteria"**. FEMS Microbiol Ecol. 2023 Mar (IF: 3.6).

9. Fontana F., Longhi G., Tarracchini C., Mancabelli L., Lugli G.A., Alessandri G., Turroni F., Milani C., Ventura M. **"The human gut microbiome of athletes: metagenomic and metabolic insights"**. Microbiome. 2023 Feb. (IF: 16).

10. Alessandri G., Fontana F., Tarracchini C., Rizzo S.M., Bianchi GM., Taurino G., Chiu M., Lugli G.A., Mancabelli L., Argentini C., Longhi G., Anzalone R., Viappiani A., Milani C., Turroni F., Bussolati O., van Sinderen D., Ventura M. **"Identification of a prototype human gut Bifidobacterium longum subsp. longum strain based on comparative and functional genomic approaches"**. Front. Microbiol. 2023 Feb. (IF: 5.6).

11. Lugli G.A., Fontana F., Tarracchini C., Mancabelli L., Milani C., Turroni F., Ventura M. **"Exploring the biodiversity of Bifidobacterium asteroides among honey bee microbiomes"**. Environ Microbiol. 2022 Dec. (IF: 5.4).

12. Fontanta F., Alessandri G., Tarracchini C., Bianchi MG., Rizzo S.M., Mancabelli L., Lugli G.A., Argentini C., Vergna L.M., Anzalone R., Longhi G., Viappiani A., Taurino G., Chiu M., Turroni F., Bussolati O., van Sinderen D., Milani C., Ventura M. **"Designation of optimal reference strains representing the infant gut bifidobacterial species through a comprehensive multi-omics approach"**. Environ Microbiol. 2022 Dec. (IF: 5.5).

13. Mancabelli L., Milani C., Fontana F., Lugli G.A., Tarracchini C., Viappiani A., Ciociola T., Ticinesi A., Nouvenne A., Meschi T., Turroni F., Ventura M. **"Untangling the link between the human gut microbiota composition and the severity of the symptoms of the COVID-19 infection"**. Environ Microbiol. 2022 Dec. (IF: 5.4).

14. Alessandri G., Fontana F., Mancabelli L., Lugli G.A., Tarracchini C., Argentini C., Longhi G., Viappiani A., Milani C., Turroni F., van Sinderen D., Ventura M. **"Exploring species-level infant gut bacterial biodiversity by meta-analysis and formulation of an optimized cultivation medium"**. NPJ Biofilms Microbiomes. 2022 Oct. (IF: 7.6).

15. Tarracchini C., Fontana F., Mancabelli L., Lugli G.A., Alessandri G., Turroni F., Ventura M., Milani C. **"Gut microbe metabolism of small molecules supports human development across the early stages of life"**. Front Microbiol. 2022 Sep. (IF: 6.0).

16. Longhi G., Lugli G.A., Mancabelli L., Alessandri G., Tarracchini C., Fontana F., Turroni F., Milani C., van Sinderen D., Ventura M. **"Tap water as a natural vehicle for microorganisms shaping the human gut microbiome"**. Environ Microbiol. 2022 Sep. (IF: 5.5).

17. Mancabelli L., Ciociola T., Lugli G.A., Tarracchini C., Fontanta F., Viappiani A., Turroni F., Ticinesi A., Meschi T., Conti S., Ventura M., Milani C. **"Guideline for the analysis of the microbial communities of the human upper airways"**. J Oral Microbiol. 2022 Jul. (IF: 9.0).

18. Argentini C., Mancabelli L., Alessandri G., Tarracchini C., Barbetti M., Carnevali L., Longhi G., Viappiani A., Anzalone R., Milani C., Sgoifo A., van Sinderen D., Ventura M., Turroni F. **"Exploring the ecological effects of naturally antibiotic-insensitive Bifidobacteria in the recovery of the resilience of the gut microbiota during and after antibiotic treatment"**. Appl Environ Microbiol. 2022 Jun. (IF: 4.8).

19. Tarracchini C., Fontana F., Lugli G.A., Mancabelli L., Alessandri G., Turroni F., Ventura M., Milani C. **"Investigation of the Ecological Link between Recurrent Microbial Human Gut Communities and Physical Activity"**. Microbiol Spectr. 2022 Apr. (IF: 9.0).

20. Alessandri G., Lugli G.A., Tarracchini C., Rizzo S.M., Argentini C., Viappiani A., Mancabelli L., Fontana F., Milani C., Turroni F., van Sinderen D., Ventura M. **"Disclosing the Genomic Diversity among Members of the Bifidobacterium Genus of Canine and Feline Origin with Respect to Those from Human"**. Appl Environ Microbiol. 2022 Apr. (IF: 5.0).

21. Mancabelli L., Milani C., Fontana F., Lugli G.A., Tarracchini C., Turroni F., van Sinderen D., Ventura M. **"Mapping bacterial diversity and metabolic**

functionality of the human respiratory tract microbiome". J Oral Microbiol. 2022 Mar. (IF: 5.8)

22. Tarracchini C., Viglioli M., Lugli G.A., Mancabelli L., Fontana F., Alessandri G., Turroni F., Ventura M., Milani C. **"The Integrated Probiotic Database: a genomic compendium of bifidobacterial health-promoting strains"**. Microbiome Res Rep. 2022 Feb.

23. Longhi G., Lugli G.A., Alessandri G., Mancabelli L., Tarracchini C., Fontana F., Turroni F., Milani C., Di Pierro F., van Sinderen D., Ventura M. **"The Probiotic Identity Card: a novel 'probiogenomics' approach to investigate probiotic supplements"**. Front Microbiol. 2022 Jan. (IF: 5.6).

24. Tarracchini C., Milani C., Longhi G., Fontana F., Mancabelli L., Pintus R., Lugli G.A., Alessandri G., Anzalone R., Viappiani A., Turroni F., Mussap M., Dessì A., Marincola F.C., Noto A., De Magistris A., Vincent M., Bernasconi S., Picaud J.C., Fanos V., Ventura M. **"Unraveling the microbiome of necrotizing enterocolitis: insights in novel microbial and metabolomic biomarkers"**. Microbiol Spectr. 2021 Oct. (IF: 7.2).

25. Fontana F., Mancabelli L., Lugli G.A., Taracchini C., Alessandri G., Longhi G., Anzalone R., Viappiani A., Famo R., Brognan M., Micondo KH., Turroni F., Ventura M., D'Alfonso R., Milani C. **"Investigating the infant gut microbiota in developing countries: worldwide metagenomic meta-analysis involving infants living in sub-urban areas of Côte d'Ivoire"**. Environ Microbiol Rep. 2021 Oct. (IF: 3.5).

26. Mancabelli L., Milani C., Anzalone R., Alessandri G., Lugli G.A., Tarracchini C., Fontana F., Turroni F., Ventura M. **"Free DNA and Metagenomics Analyses: Evaluation of Free DNA Inactivation Protocols for Shotgun Metagenomics Analysis of Human Biological Matrices"**. Front Microbiol. 2021 Oct. (IF: 5.6).

27. Tarracchini C., Milani C., Lugli G.A., Mancabelli L., Fontana F., Alessandri G., Longhi G., Anzalone R., Viappiani A., Turroni F., van Sinderen D., Ventura M. **"Phylogenomic disentangling of the Bifidobacterium longum subsp. infantis taxon"**. Microb Genom. 2021 Jul. (IF: 5.2).

28. Lugli G.A., Alessandri G., Milani C., Viappiani A., Fontana F., Tarracchini C., Mancabelli L., Argentini C., Ruiz L., Margolles A., van Sinderen D., Turroni F., Ventura M. **"Genetic insights into the dark matter of the mammalian gut**

microbiota through targeted genome reconstruction”. Environ Microbiol. 2021 Jun. (IF: 5.5).

29. Tarracchini C., Lugli G.A., Mancabelli L., Milani C., Turroni F., Ventura M. **“Assessing the Genomic Variability of Gardnerella vaginalis through Comparative Genomic Analyses: Evolutionary and Ecological Implications”**. Appl Environ Microbiol. 2020 Dec. (IF: 4.0).

30. Mancabelli L., Tarracchini C., Milani C., Lugli G.A., Fontana F., Turroni F., van Sinderen D., Ventura M. **“Vaginotypes of the human vaginal microbiome”**. Environ Microbiol. 2021 Mar. (IF: 5.5).

31. Tarracchini C., Lugli G.A., Mancabelli L., Milani C., Turroni F., Ventura M. **“Assessing the Genomic Variability of Gardnerella vaginalis through Comparative Genomic Analyses: Evolutionary and Ecological Implications”**. Appl Environ Microbiol. 2020 Dec. (IF: 4.0).

32. Lugli G.A., Tarracchini C., Alessandri G., Milani C., Mancabelli L., Turroni F., Neuzil-Bunesova V., Ruiz L., Margolles A., Ventura M. **“Decoding the Genomic Variability among Members of the Bifidobacterium dentium Species”**. Microorganisms. 2020 Nov (IF: 4.1).

33. Mancabelli L., Tarracchini C., Milani C., Lugli G.A., Fontana F., Turroni F., van Sinderen D., Ventura M. **“Multi-population cohort meta-analysis of human intestinal microbiota in early life reveals the existence of infant community state types (ICSTs)”**. Comput Struct Biotechnol J. 2020 Sep. (IF: 6.0).

34. Lugli G.A., Duranti S., Milani C., Mancabelli L., Turroni F., Alessandri G., Longhi G., Anzalone R., Viappiani A., Tarracchini C., Bernasconi S., Yonemitsu C., Bode L., Goran M.I., Ossiprandi M.C., van Sinderen D., Ventura M. **"Investigating bifidobacteria and human milk oligosaccharide composition of lactating mothers"**. FEMS Microbiol Ecol. 2020 May. (IF: 4.5).