UNIVERSITÀ DI PARMA

UNIVERSITY OF PARMA

Ph.D. in Biotechnology and Bioscience

XXXV cycle

Exploring different microbial communities through metagenomic approaches

Ph.D. Coordinator: Prof. Marco Ventura

Tutor: Prof. Marco Ventura

Industrial Tutor: Prof. Francesca Turroni

Ph.D. student:

Giulia Longhi

2019/2020 – 2021/2022

Table of Contents

Summary

The human body harbors a complex and dynamic population of microorganisms residing in various compartments such as the gastrointestinal, genitourinary and respiratory tracts and the skin surfaces, all universally recognized as the human microbiota. Among the different human-related microbial communities, the gut microbiota is certainly one of the most studied due to the high complexity and the extreme heterogeneity of the microbial ecosystems it retains, which have co-evolved over the decades to form a mutually beneficial relationship with the host from which both parties take advantages. Recently, the scientific community has turned particular interest to study the bacterial component of the intestinal microbiota since bacteria are involved in a continuous dialogue with the host affecting its health. Indeed, the gut microbiota has been found to be crucial for immunologic, hormonal, and metabolic homeostasis of the host. However, several factors, like the environment and the host's lifestyles, may cause modifications in the microbiota composition, thus causing dysbiosis, which is often related to the onset of many diseases. Traditionally, the assessment of these bacterial communities has been based on conventional culture-dependent methods that do not allow an exhaustive characterization of the microbial biodiversity occurring in the human body. In the last decades, the advent and development of culture-independent approaches based on Next-Generation Sequencing (NGS) techniques, also known as metagenomics, allowed an accurate disentangling of the microbial populations inhabiting the human body. Specifically, metagenomics attempts offer the possibility to profile the bacterial taxonomy and predict the functional activities exploited by the microbial communities and thus underpinning the microbe-microbe and microbe-host interactions. Thanks to NGS approaches, the interest in studying the microbiome and its role in the establishment and maintenance of human health, also for diagnostic purposes has strikingly increased. Despite that, the study of microbiome

composition in human biological specimens does not always come without challenges due to many procedural issues.

The aim of this Ph.D. thesis is to explore the composition of different microbial communities by means of the most reliable metagenomics approach. Specifically, it aims to investigate the reliability of a widely used protocol for the depletion of eukaryotic DNA through the analysis of different human specimens rich in host DNA, focusing on the impact of this protocol on the detection of bacterial populations.

Furthermore, through NGS techniques, the study of the microbiota, with a particular interest in the gut microbiota, also highlighted aspects concerning its modulation. In this context, there is growing scientific evidence on the correlation between the alteration of gut microbiota composition and disorders. Many potential factors such as diet, prebiotic compounds and probiotic products can modulate the human intestinal microbiota, trying to re-establish an altered bacterial composition. However, despite the wide commercial employment of probiotic formulations, very little is known about the molecular mechanisms of the action and the genetic features of probiotic bacteria. In this context, this Ph.D. thesis aims to unravel the microbiota composition of different probiotic products present on the Italian market by developing a powerful and reliable pipeline combining whole metagenome shotgun analyses and flow cytometry assays to verify the quality of probiotic formulations.

Moreover, among the various factors that can influence the gut microbiota composition, there is solid and liquid diet. Concerning the gut modulation acted by food, in this Ph.D. thesis, we investigated the resident microbiome of many Italian raw milk cheeses being part of Protected Designation of Origin (PDO) denomination, revealing how the microbiota harbored by each cheese is mainly linked to the type of cheesemaking process together with local environmental factors

rather than exclusively to the cheese type or the geographical origin. The study of food microbiota represents the first necessary step to get an overview of the possible influence these microbial communities could exploit on the consumer's intestinal microbiota. Finally, water is in all respects considered as a food but also as a reservoir of microorganisms able to colonize and modulate the consumer's intestinal microbiota. In this context, one of the purposes of this Ph.D. thesis is to evaluate the microbial composition inhabiting drinking water through a comprehensive shotgun metagenomics analysis of tap water microbiome, highlighting the occurrence of a deeply bacterial biodiversity and the presence of a conserved core tap water microbiota most represented by unknown microbial species, constituting the so-called microbial dark matter. Furthermore, genome reconstruction of the dominant bacterial genera of water microbiota allowed us to unveil their presence in the fecal microbiome of humans from various geographical locations, providing evidence of a potential novel route of horizontal microbial transmission.

Chapter 1

General Introduction

A. The human microbiota

Human body encompasses a complex and dynamic population of microorganisms, collectively referred to as the microbiota (1). The microbiota includes not only the bacterial communities colonizing a specific environment, but more broadly encompass the set of fungi, Archaea, viruses and protozoans that populate this ecosystem (2). The human body includes several ecological niches/microenvironments due to different chemical and physical conditions encompassed in this complex environment. So far, the most investigated human body compartments with respect the microbial communities include the gastrointestinal tract (GIT), the vaginal tract and the skin areas. Large part of microbes resides in the GIT, particularly in the distal intestine, and are estimated to exceed approximately 10 times the total number of human somatic and germline cells (3). The gut microbiota can be considered as an additional organ of the human body in which a considerable number of microorganisms are able to communicate with each other and with the host. Some of these microorganisms are natural inhabitants of the GIT like autochthonous bacteria, while others are transient or allochthonous microorganisms able to reach the intestine by ingestion of food and by possible environmental contaminations (4). It has been postulated that the bacterial taxa encompassing the intestinal microbiota involve more than 500 species (5). In this context, the most abundant species consist in the Firmicutes (51%) and Bacteroidetes (48%) phyla, while the bacterial species less abundant belong to Actinobacteria, Proteobacteria and Verrucomicrobia phyla (6, 7). This leads not only to a different number of microbial species in our body, but also a different distribution of these bacterial cells in our GIT.

During the evolution of human beings, the gut microbiota has established a mutualistic symbiosis with the host. In fact, the host provides the nourishment for gut microbes and a suitable environment for the growth of the microbial community, which supplies, through metabolism,

essential products for the establishment and maintenance of the host's health (1, 8-10). There are many functions associated with the intestinal microbiota that are fundamental for maintaining a correct physiology of the host. For example, the intestinal microbiota produces various glycosyl hydrolases allowing the metabolism of indigestible polysaccharides and subsequently making these accessible to the host's metabolisms. In addition, the gut microbiota promotes the biosynthesis of essential amino acids and vitamins, the removal of toxic compounds, the production of energy for intestinal epithelial cells and strengthens the mucosal barrier through the synthesis of short-chain fatty acids (SCFAs) such as propionate, butyrate, and acetate (11, 12). The gut microbiota can also promote epithelial cell maturation and angiogenesis. It has also been found to be essential in guiding adaptive immunity in recognizing and responding to specific microorganisms or programming many aspects of T-cell differentiation. In this context, many studies involving germ-free mice (mice born and raised in the absence of any microorganisms) provide essential insights into how intestinal epithelial cells show reduced expression of molecules involved in pathogen sensing and antigen exposure (13, 14).

B. Methodologies for the cataloguing of the human microbiota

Microorganisms are abundant and ubiquitous; however, we still lack a fundamental mechanistic understanding of many of the key roles played by these microorganisms in nature, including those supporting their colonization of the human body. Thus, it is of crucial importance to assess the composition of the human microbiota through a detailed cataloguing of the various microbes constituting these microbial communities. Up to now, there are two main strategies aimed at studying the composition of the microbiota: i) the culture-independent approaches, ii) the culture-dependent methods. The formers include classical microbiology techniques, involving the use of selective culture media, allowing the growth and subsequent isolation of bacterial cells and their subsequent characterization through metabolic and physiological assays (15). However, these approaches display serious limitations of allowing an appreciable detection only of the cultivable fraction of the bacterial community, omitting instead the large fraction of microorganisms that cannot be cultivated under laboratory conditions due to their very particular nutritional requirements. It is worth mentioning that such uncultivable bacteria in some cases represent 70% of the total bacterial population, thus representing the largest proportion of the bacterial biodiversity. Recently, such proportion of uncultivable bacteria has been denominated as the microbial dark matter and their understanding represents an important new frontier of microbiology research applied to the study of microbiomes (16-18). Culture-independent approaches, on the other hand, through metagenomic analyses, consent a more accurate study and characterization of the bacterial communities that compose the intestinal microbiota, allowing to access also the latter non-culturable fraction through the assessment of their genetic repertoire (19).

In this context, high-throughput sequencing of the 16S rRNA gene as a conserved phylogenetic marker has been considered for a long time as the golden-standard technique for profiling complex microbial communities, but nowadays shotgun metagenomics is gradually prevailing. The 16S rRNA gene-based microbial profiling approach leverages universal primers for the amplification of hypervariable regions of the 16S rRNA gene (20). The amplicon reads obtained through Next Generation Sequencing (NGS) platforms, once processed through bioinformatics pipelines, generate a taxonomic profile of the analyzed samples and thanks to their comparison with the current existing 16S rRNA gene datasets, allows the identification of unknown bacterial members through the discriminative power of their unique hypervariable regions (20, 21). Therefore, 16S rRNA microbial profiling analysis has always been considered a robust and well-characterized method capable of providing sufficient information on the composition of microbial communities, starting from a relatively small number of reads per sample. However, one important limitation of this approach is that microbial taxa are assigned based on the DNA sequencing of only one region of the bacterial genome (22), thus lacking an appropriate taxonomic resolution that is limited down to the genus level (23). Conversely, shotgun metagenomics consists of sequencing the total bacterial genomic DNA isolated from the analyzed sample (24). Although it requires a higher coverage (10-30 million reads) that requested by 16S rRNA microbial profiling assays and a more complex data analysis, it allows a much deeper characterization of the complex microbiome and more accurate and precise identification to species/strain level than 16S rRNA amplicon sequencing (25).

However, to obtain information on the functions exploited by the microorganisms within complex bacterial communities, at a given moment and in certain environmental conditions, it is necessary to apply functional metagenomic approaches such as metatranscriptomics, which consists in the

sequencing of an entire pool of microbial RNA (26, 27), or metaproteomics and metabolomics that mapped the set of proteins/peptides produced by mixed bacterial communities, providing functional information and allowing to monitor modifications in protein expression within the microbiota in response to changes in normal environmental conditions (28, 29). As for metagenomics, also for these latter functional approaches, there are several technical limitations, for example many genes as well as their products have not yet been fully characterized.

Culture-independent approaches for the investigation of the human gut microbiota composition

Several metagenomic techniques have been developed and applied in the last decade to study the gut microbiota composition, such as the 16S rRNA microbial profiling as well as the internally transcribed spacer (ITS) profiling, and shotgun metagenomics (Fig. 1).

The 16S rRNA microbial profiling approach exploits the ribosomal 16S rRNA gene, which together with the 23S gene, the 5S gene and the ITS region constitutes the bacterial ribosomal operon (30). The ribosomal 16S-rRNA gene has a length of about 1500 nt and it is characterized by the presence of nine hypervariable regions called V1-V9, alternating with as many highly conserved RNA loci and weakly subject to mutational events. In particular, the hypervariable subunits V3, V4, V5 and V6 are exploited for the phylogenetic analyses, allowing a taxonomic resolution down to genus level. These regions are commonly applied in 16S rRNA microbial profiling analyses of the complex microbial communities present in different ecosystems, including the different districts of the human body (21, 31).

To obtain a deeper snapshot of the human gut microbiota composition at the species or even subspecies level, it is necessary to target a much more variable molecular marker than the 16S rRNA gene. A valuable genetic marker for such a purpose is represented by the Internally Transcribed Sequences (ITS), which is a spacer region located between the 16S and 23S genes of the bacterial operon, characterized by the fact that it displays a high degree of nucleotide diversity even among closely related species. This implies the possibility of its use in phylogenetic studies among bacterial populations, since through ITS microbial profiling it is possible to classify

bacteria up to the species and subspecies level, instead of 16S rRNA profiling, which as above mentioned, is restricted to a genus level classification (32, 33). Remarkably, ITS region has been successfully applied for the detailed profiling of specific microbial groups commonly residing in the human gut such as bifidobacteria and lactobacilli (34). Furthermore, the ITS region has also been used for the identification of species present in numerous fermented and functional foods, and in the vaginal microbiota (35).

Recently, a new bacterial profiling method based on the ITS region has been developed employing universal ITS primers for bacteria, which were designed through the alignment of 16S-ITS-23S sequences, and an inclusive ITS database for the accurate classification of bacterial communities at (sub)species level (36). The collected data showed that this novel metagenomic pipeline represents a breakthrough in the identification and screening of microbial taxa at the (sub)species level, with significant relevance not only in research but also in the industrial and clinical field (36).

Another metagenomic approach applied to investigate the gut microbiota composition at a very detailed taxonomic resolution, i.e., down to the strain level, is represented by shotgun metagenomic approach that in some cases allows the reconstruction of the genome sequences of the microbes present in the biological sample analyzed. The shotgun metagenomics performs the sequencing of the total genomic DNA present in the ecosystem and therefore might lead to access to the complete genetic makeup of the various microbes present in a given sample. It is a technique allowing a broader taxonomic and functional classification compared to the other above described metagenomic approaches, thus leading to the possibility of making *in silico*-based predictions relating to the metabolic capacities of a bacterial community (37, 38). In fact, in contrast to 16S rRNA and ITS microbial profiling analysis approaches that allow a predictive characterization of

the functional capabilities of a certain bacterial species only through PCRUSt (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States), the shotgun metagenomics consents the study of these characteristics in a direct way, also evaluating the variations that a microbial ecosystem as the gut microbiota may be subjected over time in response to a given change (39).

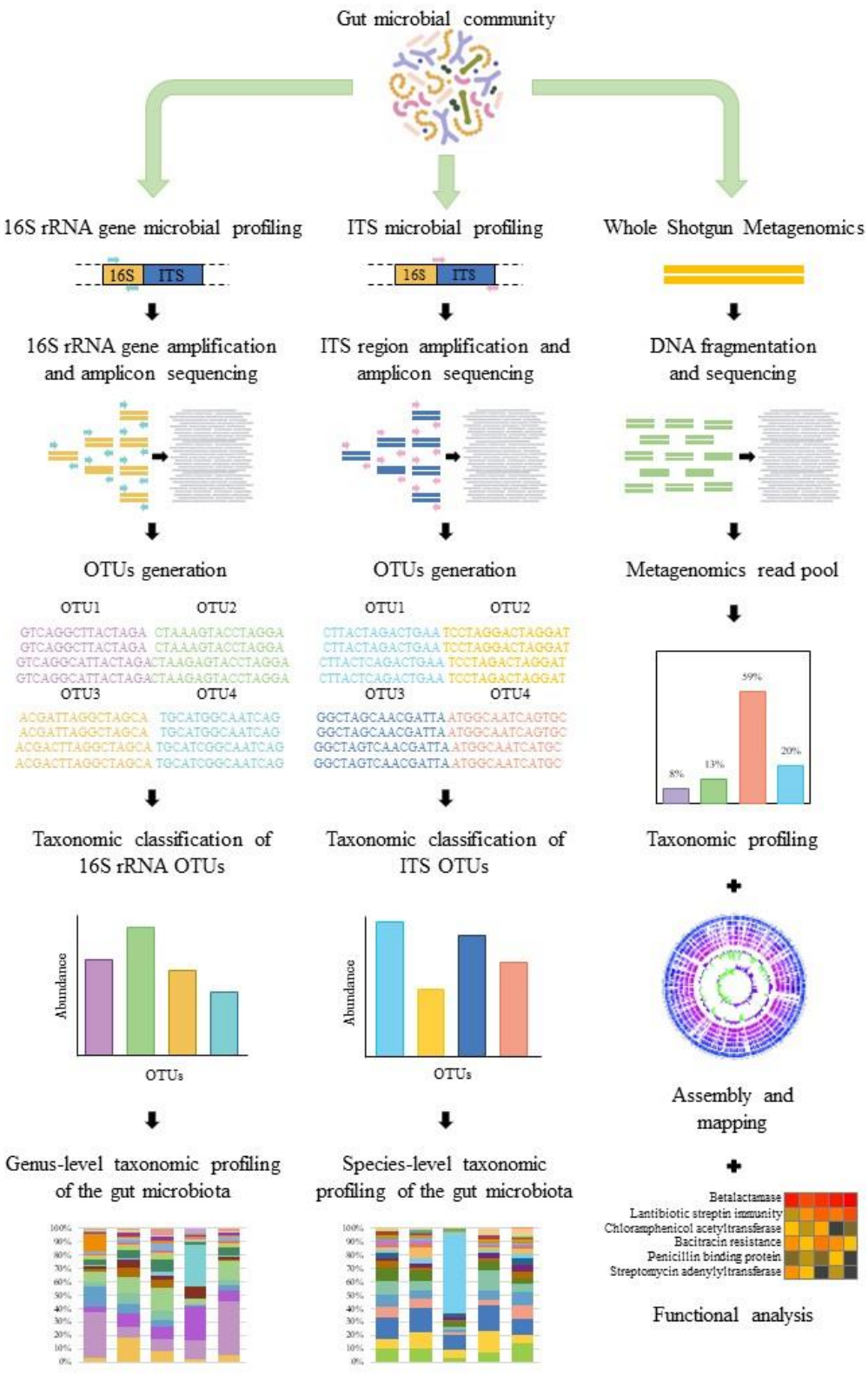


Figure 1: Metagenomic approaches for the microbiota characterization [modified by (17)].

Limitations of the targeted sequencing approach in microbiota analysis

Polymerase chain reaction (PCR) amplification is a fundamental step in the profiling of microbial communities through 16S rRNA gene and ITS profiling approaches (40). Nevertheless, the different amplification efficiencies could sometimes produce bias, which can hamper the correct evaluation of the community structure, representing a source of error for the microbiota studies (41-43). PCR-induced artifacts could occur at any distinct step of amplification and could be associated to i) sequence artifacts resulting from undesired PCR products; ii) PCR bias altering the distribution of PCR products due to unequal amplification or cloning efficiency (42). Sequence artifacts may be due to Taq DNA polymerase errors (44), primers mismatch in the first PCR cycles (45), the formation of heteroduplex molecules (46) or chimerical molecules (47). However, it is generally thought that PCR bias is probably linked to intrinsic differences in the amplification efficiency of templates (48) or to the inhibition of amplification caused by the self-annealing of the most abundant templates during the last stages of the PCR reaction (43). Although several approaches have been proposed to optimize the amplicon-based method, including the reduction of the number of PCR cycles (49), primers optimization (50) and polymerase optimization (51, 52), the PCR bias remains a very delicate step of metagenomic protocols and represents an important source of error in microbiome studies.

Another possible delicate step in the targeted sequencing approaches includes the DNA sequencing and amplification fault, which generate DNA sequences that are often difficult to identify (53). In fact, if these are applied to still unknown microbes, i.e., not yet present in the publicly available repositories, their identification would be very problematic (54).

C. Impact of the DNA extraction procedures on the assessment of the microbial community composition

As previously mentioned, shotgun metagenomic sequencing is progressively replacing the 16S rRNA gene microbial profiling technique, allowing to decode all the microbial genomic DNA extracted from the analyzed environmental sample, including not only bacterial DNA but also DNA from viruses and other eukaryotic microorganisms.

However, when shotgun metagenomics is applied, it is indispensable to avoid any contamination with the host's DNA (55). This implies the need to follow a very careful and scientifically validated protocol, including biological sample collection, microbial DNA extraction, DNA sequencing, and *in silico* analysis of the data.

A critical point in the sample preparation pipeline is represented by the extraction of the microbial DNA that should be highly representative of the real microbial biodiversity of the biological samples assessed. The risk in processing biological samples highly contaminated by eukaryotic DNA is to obtain at the end of the DNA sequencing step many eukaryotic reads. Thus, a critical step in a valuable shotgun metagenomics workflow is represented by the removal of host DNA from the biological sample. In this context, the most often used approaches to eliminate host-derived DNA include those acting before or during DNA extraction and those acting after DNA extraction (56).

Depending on the biological matrix as well as the laboratory methodologies applied, there are several challenges associated with successful microbiota analysis (57). Concerning human biological matrices, due to the low bacterial load characterizing some human body compartments

(e.g., mucosal biopsies, saliva, and nasopharyngeal samples), microbial DNA extraction is biased by a large amount of human DNA carryover. In this context, novel approaches have been developed to enrich bacterial DNA from the biological sample. However, they have been proven to be partially ineffective (58, 59).

Conversely, another approach is aimed at depleting host DNA from the biological sample. Many commercially available DNA extraction kits frequently applied in microbiome studies (55, 60, 61) and the application of different methods including solvents, detergents like saponin combined with DNase (62, 63), tween 20, triton X-100 (59) or benzonase (64), have been proposed as useful expedients to remove host DNA without affecting the integrity of the microbial DNA. All these techniques are based on the different susceptibility of cells to lysis and succeeded specifically in pathogen detection, despite a high microbial load and a large sample volume often required (65, 66).

Although many studies have highlighted the valuable outcomes of these approaches, evidence is emerging showing how these techniques are somehow influencing microbial DNA extraction and thus ultimately are impacting the final determination of the gut microbiota composition.

In this framework, post-extraction approaches try to avoid some of the difficulties linked with microbial DNA extraction protocols and thus represent valuable alternatives. Bacterial DNA enrichment commonly involves hybridizing specific probes to the target bacterial genome or host-specific sequence probes for CRISPR/Cas9 hybridization-based depletion (67). However, these post-extraction microbial DNA enrichment approaches require very high costs, even for processing small sample sets.

Given the importance of the microbiota and specifically of the gut microbiota in human health, there is the need to have a well-developed, standardized, and affordable protocol in order to

achieve a reliable investigation of the gut microbiota composition, contributing to microbiome research.

D. Origin of the human gut microbiota

The development of the human intestinal microbiota represents a process in which positive and negative interactions between different microbial taxa occur after delivery and is influenced by different perinatal conditions, such as the type of birth and the weaning method (2, 68) (Fig. 2).

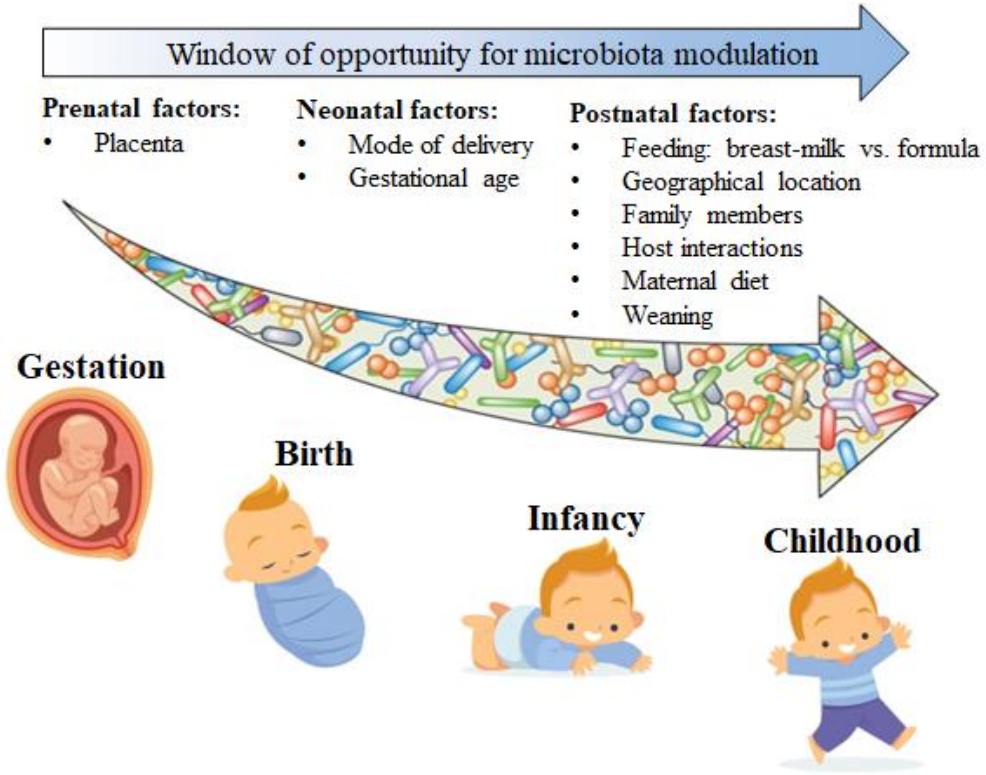


Figure 2: Potential factors modulating microbiota settlement and development from gestation to infancy [modified by (69)].

Furthermore, it is assumed that factors not directly implicated in the newborn's development during the first months of life may be involved, like diet, family genetic components and environmental factors (70). It was a common dogma in the scientific community that the uterine cavity, during the gestational period, is a completely sterile environment (71). However, some recent studies have

reported possible signs underlying a microbial colonization of the fetus that is influenced by the maternal microbiota (72, 73). Until now, it has been shown that the delivery method represents the factor exploiting the larger influence on the establishment of the human intestinal microbiota (74). Specifically, infants born with natural delivery possess an intestinal microbiota that is expected to be maternally inherited through a vertical transmission by the birth canal. The gut microbiota of vaginally delivered infants is influenced by microorganisms inhabiting the maternal urogenital tract, with a high prevalence of bacteria belonging to the genus *Lactobacillus* and *Prevotella* (75, 76). On the contrary, babies born by Caesarean section are mainly colonized by bacteria deriving from the external environment, therefore by microorganisms horizontally transmitted through the mother's skin and the hospital environment with a prevalence of *Staphylococcus*, *Corynebacteria* e *Propionibacterium* spp. (77-81). Another important factor largely influencing the composition of the infant gut microbiota is represented by the diet. Breastfeeding in the first months of life is considered to exploit an important role in the settlement of the intestinal microbiota of the newborn (82). In this context, numerous differences related to the type of infant feeding have been found. Specifically, breastfed infants showed high levels of lactobacilli and bifidobacteria, which are considered as valuable microbial biomarkers indicative of a healthy human host (82, 83). Breast milk represents an essential source of nourishment for the newborn and includes numerous beneficial components for human health. It is a fluid formed by a mixture of lipids, vitamins, proteins, lactose, and oligosaccharides known as Human Milk Oligosaccharides (HMO) and essential factors for the infant's immune system like immunoglobulin A (70). HMOs are, together with lactose, the most abundant carbohydrate component of breast milk and stimulate the colonization of members of the *Bifidobacterium* genus during the early stages of the host's life.

Notably, HMOs are not metabolized by mammalian enzymes, which in contrast, are typically hydrolyzed by few species of the *Bifidobacterium* genus.

On the other hand, infants weaned with formula milk display a higher abundance of bacteria belonging to genus *Bacteroides*, *Clostridia*, *Staphylococcus* and the Enterobacteriaceae family. Thus, representing an infant microbiota more complex than the one observed in breastfed infants and possessing many similarities to that of an adult individual (82, 84, 85). These differences may have important implications on the host physiology and immunology, contributing to disease susceptibility including asthma and atopic diseases in adulthood (86, 87).

Subsequently, the first colonization of the human GIT is subjected to major changes until it reaches maturation after the weaning. Noticeably, only a subset of microbes to which infants are initially exposed is expected to permanently colonize the human gut microbiota. After the third year of life, the human gut microbiota becomes relatively stable, assuming a certain complexity that makes it like that encountered in the large intestine of the adult (70, 88). It has been established that in the first months of life, the intestinal microbiota is mainly dominated by microorganisms belonging to Actinobacteria and Proteobacteria phyla, then over time it has been found an increase in colonization by bacteria of Bacteroidetes and Firmicutes phyla (70, 89).

E. Factors involved in the modulation of the composition of the human gut microbiota

As previously mentioned, the investigation of the gut microbiota has gained remarkable interest since then the gut microbiota has been shown to play a crucial role in the establishment and maintenance of the host's health during the entire life span of human beings. The consolidation of a microbial ecosystem within the GIT is a process that takes place in the newborn; it is formed "*de novo*" and it is consequently submitted to various changes during the first months of an individual's life, being influenced by endogenous and exogenous factors (90, 91). Such perturbations in its composition during development and settlement have been linked to pediatric disorders and disease's onset in adulthood (92).

The gut microbiota plays multiple roles in promoting the host's health such as stimulating the immune system, protecting against pathogens, or promoting the differentiation of host cells. However, also the host exploits an important role in the establishment of the human gut microbiota. In fact, through a set of resources and conditions, the human gut promotes and allows the specific growth of some microorganisms (70). Moreover, the gut microbiota provides several compounds, like lipids and proteins, able to interact with host epithelial cells. Specifically, epithelial cells have developed separate mechanisms dedicated to detecting the presence of molecules such as lipopolysaccharide (LPS) and peptidoglycan, which represent the main component of the microbial cell surfaces.

The composition of the gut microbiota could also be influenced by exogenous natural factors like aging, drugs, physical activity versus a more sedentary lifestyle, and cultural or geographical locations. In this context, the different geographical origins of everyone could be responsible for

the disappearance of low-abundant species such as antibiotic-sensitive strains or the acquisition of others following industrialization and urbanization (93).

However, the gut microbiota is a very dynamic ecosystem in which it is difficult to identify a stable conformation corresponding to a standardized composition corresponding to a "healthy" microbiota. The term "enterotypes" has been recently introduced to identify three distinct well-balanced host-microbial symbiotic states assumed by the gut microbiota that might respond differently to diet and drug intake (94). These three enterotypes are defined by the abundance of different genera such as *Bacteroides*, characterizing Enterotype I, or *Prevotella* and *Ruminococcus* specifically distinguishing Enterotypes II and III, respectively (94). Based on these models, enterotypes can provide important insights into the functions performed by their members such as the *Bacteroides* genus is frequently associated with a diet rich in animal proteins and saturated fats, while *Prevotella* genus is commonly associated with a high-fiber diet (94).

The human gut microbiome classification in the form of distinct community composition has been welcomed with great interest and controversy. The original definition clarified that enterotype classification is an easy way to define and stratify samples reducing their complexity (95), although this stratification has often been used for the analysis of microbiome data (95). Enterotype classification showed many limitations related to sampling and selection bias or to the different culture-independent methods and clustering applied (96). Finally, the numerous analyses performed have produced excessive confidence in the determination of enterotypes without considering that there is a continuous modification of gut microbiota composition even within the same subject (96-98).

F. Correlation between the human gut microbiota composition and disease

As previously mentioned, the microbiota is composed by microorganisms trying to live in a balanced equilibrium with the host, a status known as eubiosis, but several factors such as the environment and the lifestyle of the host, may cause shifts in the composition of the microbiota, causing dysbiosis (99) (Fig. 3).

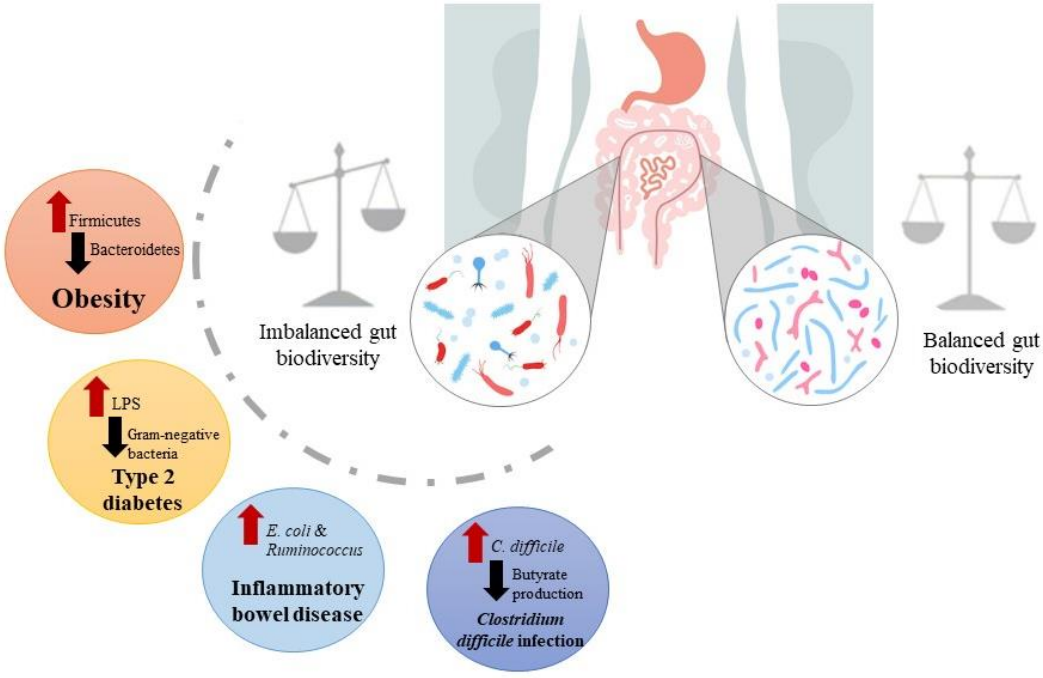


Figure 3: Schematic illustration of the factors causing shifts in human microbiota composition.

Gut dysbiosis exploits a negative impact on host health with long-term health consequences, being related to several diseases or disorders (70). Dysbiotic alterations of the microbiota generally involve the loss of health-associated microorganisms (mainly SCFA producers), with the increase of opportunistic pathogens such as mucolytic bacteria (producers of hydrogen, methane, and

hydrogen sulfide) proteobacteria with an increase in LPS endotoxin. These might produce negative consequences on the health status of the host, in terms of compromised integrity of the intestinal mucosa, acute inflammation of the mucosa itself with translocation of bacteria, toxic effects on colonocytes, oxidative damage, alteration of the cytokines pattern and other systemic effects (100-102).

Many human gut microbiota-based studies (103, 104) have produced crucial information on the structure and functional potential of the microbial communities populating the human intestine, highlighting a high interindividual variability and an apparent intraindividual stability (105). These studies also allow differentiating conditions of eubiosis from dysbiosis.

Specifically, it was possible to characterize the dysbiotic profiles of the microbiota in the context of gut pathologies such as Crohn's disease, ulcerative colitis, non-alcohol-related liver dysfunction and disorders like irritable bowel syndrome (IBS), obesity and type 1 and 2 diabetes (106-108). Furthermore, potential relationships between an altered structure of the intestinal microbiota and other pathological profiles are now being investigated, including autism spectrum disorders, multiples sclerosis and neurodegenerative diseases, as well as oncological diseases (109-111). However, recent meta-analyses suggested that intestinal microbial dysbiosis can change over time and the characterization of their dynamics is important in defining personalized therapeutic strategies (112, 113).

An important index of the possible altered eubiosis status is also represented by the decrease in microbial diversity, which is evaluated through α-diversity analysis based on the quantitative evaluation of the number of bacterial taxa present in a community. In dysbiosis condition, a lower diversity is usually characterized by the loss of the most abundant commensal taxa (114, 115). Notably, a decrease in richness of the microbiota population is described as a biomarker for

metabolic disorders, but it is not sufficient to confirm the presence of an ongoing pathological state (70).

Microbial biomarkers associated with the human health

There is growing interest in understanding the precise molecular mechanisms supporting gut microbiota alterations because they could be utilized for novel diagnostic and prognostic assays. In this context, bifidobacteria are important commensals of the human gut, especially in the first stages of life (116), but their abundance remains relatively stable in adulthood. The presence and abundance of bifidobacteria are very often associated with a healthy status of the host and for this reason they can be considered as valuable novel microbial biomarkers (70). For example, low levels of bifidobacteria during childhood have been positively correlated with celiac disease, obesity, autoimmune diseases and colic and necrotizing enterocolitis (NEC) (117, 118). In the latter case, a recent study demonstrated that intestinal dysbiosis could lead to the onset of NEC, showing an overgrowth of opportunistic microbial species accompanying the loss of gut microbial biodiversity and suggesting the involvement of *Clostridium* genus members as potential predictive biomarkers for early diagnosis of this disease (119). Other important microbial biomarkers are members of the *Lactobacillus* genus colonizing different ecological niches of the human body (120). One of the main niches is represented by the female genital tract and the most common species here identified encompass Lactobacillus *crispatus*, *Lactobacillus gasseri* and *Lactobacillus jensenii*. A microbiota dominated by lactobacilli appears to represent a suitable microbial biomarker for the vaginal health of adult women because this genus can establish a barrier against pathogen invasion by producing antimicrobial metabolites preventing bacterial and viral infections (121).

Moreover, individuals with inflammatory bowel disease (IBD) show deficiencies in *Faecalibacterium prausnitzii* and *Roseburia* spp., while presenting a large abundance of

Firmicutes and Enterobacteriaceae (122, 123). *F. prausnitzii* is one of the major commensal bacteria of the intestinal microbiota and its presence is an index of good health. Although the role of this microorganism is still being studied, its anti-inflammatory activity, stimulating the production of cytokines such as IL-10, has been widely demonstrated (124, 125). Conversely, *Bacteroides fragilis* has been proposed as a microbial biomarker for IBD development due to the production of an enterotoxin (*B. fragilis* toxin or BFT) that causes diarrhea and inflammation in the intestine possibly resulting in the onset of colorectal cancer (126, 127).

An additional microbial biomarker associated with human disease is represented by *Fusobacterium nucleatum*, which is a typical colonizer of the oral cavity and seems to play a crucial role in dental caries, where its abundance is related to its ability to collaborate with other microorganisms of the buccal cavity (128, 129). Therefore, it is recognized as a potentially pathogenic microorganism for humans. Recently, it has been proposed as a biomarker for colon cancer in preclinical and clinical studies (130-132).

Regarding microbial biomarkers associated with host health in the first stages of life, a recent analysis showed that high levels of *B. fragilis* at one month of age are significantly correlated with higher body mass index in infants, resulting in obesity and obesity-related disorders (70). Moreover, studies in children with type 1 diabetes showed a lower diversity and significant differences in the ratios of Firmicutes and Bacteroidetes and an equivalent reduction in the abundance of the butyrate producer *F. prausnitzii* (115).

The identification of possible early signals of alteration from eubiosis is highly desirable for the establishment of novel preventive approach.

G. Prebiotics as modulators of the gut microbiota composition

For the entire lifespan of the host, the intestinal microbiota undergoes changes, and such shifts have been proven to have important consequences for establishing diseases and disorders (133). To overcome this problem, new strategies have been developed to promote the re-establishment of the eubiosis of the human gut microbiota, mainly by using nutraceutical products based on prebiotics and/or probiotic bacteria.

Prebiotics are defined as non-digestible and fermentable food components capable of determining a selective stimulation of the growth and the activity of one or more potentially beneficial commensal bacteria of the intestinal microbiota provided with beneficial properties for the host (134). Prebiotics allow the increase not only of the number of beneficial bacteria already present in the colon but also of their metabolic activities through the supply of fermentable substrates with a reduction of potentially pathogenic microorganisms belonging to Clostridiaceae and Enterobacteriaceae taxa (135, 136). The first prebiotics to be described are complex plant carbohydrates like inulin-type fructans (137). Inulin is a non-digestible oligosaccharide that, for nutritional labeling, is classified as dietary fiber and it is naturally abundant in various vegetables and fruits like agave, artichokes, asparagus, bananas, chicory root, garlic, onions, leeks, and wheat (138, 139). Several studies have investigated the impact of inulin on the human gut microbiota. They have revealed that large part of the gut microbiota remains unaffected by the presence of these glycans. At the same time, only a small number of taxa increased in their abundance, such as *Bifidobacterium* (140). Nevertheless, it is partly unclear how the level of bifidobacteria increases as they cannot degrade inulin (141). It is possible that bifidobacteria can establish trophic interactions with other gut commensals through the cross-feeding driven by other members of the

gut possessing the extracellular enzymes necessary for the breakdown of inulin, allowing the release of products (e.g., fructooligosaccharides) that will represent an important carbon source for bifidobacteria (142). Furthermore, recent research has demonstrated that inulin increases the levels of some bifidobacterial species thanks to the action of the β-fructofuranosidase that catalyzes the degradation of both sucrose and inulin (143). Other natural compounds to be identified as prebiotics were also those contained in breast milk like HMOs. HMOs are indeed considered natural prebiotic compounds because they actively stimulate the growth of the specific members of the newborn's gut microbiota. In this regard, HMOs show the ability to specifically promote the growth of some bifidobacterial taxa that successfully metabolize HMOs (144). The prebiotic effect elicited by HMOs is due to their peculiar chemical structure; in fact, some bacteria metabolize type 1 structures (terminal Gal1-3GlcNAc), while others prefer type 2 structures (Gal1-4GlcNAc) or branched structures. Specifically, *Bifidobacterium longum* subsp. *infantis* is able to hydrolyze the core structure (lacto- N-tetraose [LNT] or lacto-N-neo-tetraose [LNnT]) of HMOs (145).

Due to their chemical structures and the consequent inability of the host to digest them, prebiotics are directly metabolized in the colon by endogenous bacteria, which very often can produce SCFA, with a consequent reduction in pH. Thanks to this process, they can exert anti-inflammatory effects, for example by stimulating the increase of regulatory T cells and the reduction of interferons (146). As for HMOs, there are multiple mechanisms by which they help shape microbial communities in the infant gut, such as exploiting antimicrobial effects against certain bacterial groups like group B *Streptococcus* (GBS) and stopping growth in the presence of HMOs (147). Moreover, it has been proved that HMOs appear to impact yeast cells, altering the morphology and length of the hyphae and their attachment to epithelial cells (148).

Other chemical compounds studied as prebiotics were those capable of promoting the growth of lactic acid-producing microorganisms, such as lactulose, which is frequently used in infant formulas to increase the number of lactobacilli in the gut and to promote the development of a neonatal gut microbiota very close to that of breastfed infants (149-151). In this context, the main prebiotics used are galacto-oligosaccharides (GOS), fructo-oligosaccharides (FOS) and polydextrose (PDX). Many studies support the notion that adding GOS or a GOS/FOS mix to infant formula positively affects lactobacilli and bifidobacteria abundance by improving digestive and immune health (150, 152, 153). In detail, infants fed with GOS showed an increase in *Bifidobacterium breve* and a decrease in *Clostridium difficile* compared to infants fed with formula (154). In addition, preterm-born infants fed with formulas enriched with FOS showed an increased abundance of bifidobacteria in their fecal samples and a reduction of *Escherichia coli* and enterococci (155).

Based on this evidence, prebiotics represent a good strategy to modulate intestinal microbiota composition and prevent disease onset. However, further studies are needed to highlight the functional differences between types of prebiotics and their combination.

H. Probiotic bacteria and their influence on the human health

The modulation of the gut microbiota using probiotic bacteria is currently a promising approach in the prevention of numerous diseases. In 2001, the Food and Agriculture Organization of the United Nations (FAO) and the World Health Organization (WHO) defined probiotics as "live microorganisms which when administered in adequate amounts confer a health benefit on the host" (156, 157). However, the concept of probiotics dates to 1908 when Metchnikoff noticed that the consumption of some fermented foods led to positive effects on human health (158, 159). Probiotic bacteria are widely used in preventing various disorders affecting GIT by interacting on several levels in the regulation of the gastrointestinal barrier and thus promoting the re-establishment of the eubiosis status (160) (Fig. 4).

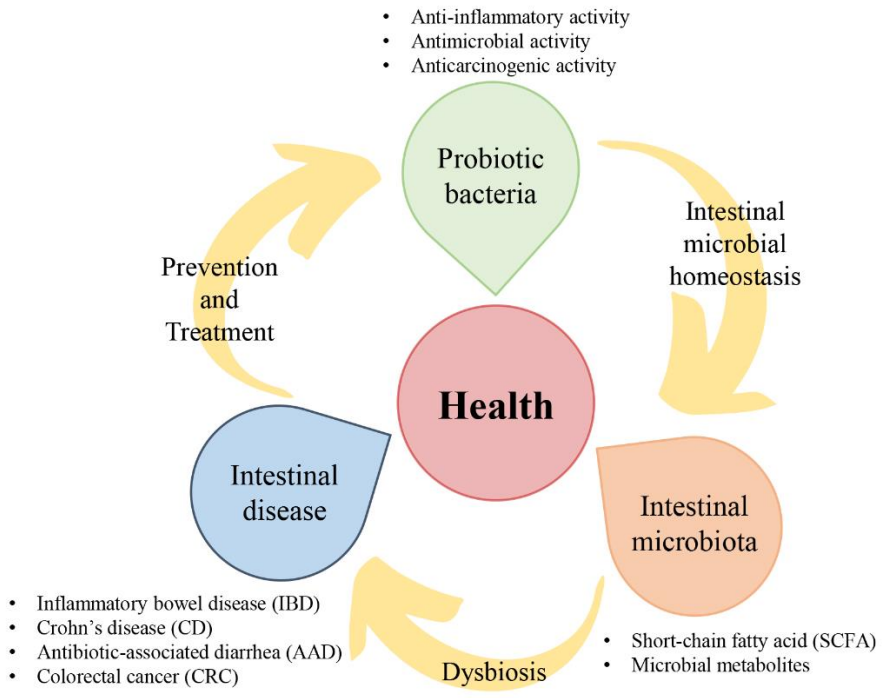


Figure 4: Probiotic bacteria effects on the human gut microbiome.

The microorganisms commonly exploited as probiotics are both bacteria and yeast, belonging to *Lactobacillus*, *Bifidobacterium*, *Enterococcus*, *Bacillus*, *Escherichia,* and *Saccharomyces* genera (146). According to the Italian Ministry of Health guidelines, a probiotic microorganism must fulfill specific criteria to be used in foods and probiotic supplements, and above all, it must be considered safe. This implies that probiotic products must satisfy precise genetic requisites, such as univocal taxonomic identification and precise evaluation of their genetic content (161), thus fulfilling criteria of safety and functionality, as well as technological utility (162). In this regard, the safety of a probiotic microorganism is defined by the lack of any characteristics of pathogenic microorganisms in terms of producing or favoring the metabolism of toxic substances and by its antibiotic resistance profile. Furthermore, to perform their beneficial effect on human health, probiotic bacteria must be able to survive in the various physiological conditions of the GIT, resisting bile salts and oxidative stress in the large intestine and to the exposure to low gastric pH (163), and they must show the ability to adhere to the intestinal mucosa (158). To satisfy all these requirements, a probiotic bacterium candidate should originate from humans and thus should belong to the autochthonous human microbiota.

There are many mechanisms by which probiotic bacteria are supposed to exert their health-promoting effects. In addition to modulating the microbiota composition, they should display the ability to increase the production of mucin by the goblet cells, to strengthen the apical tight junctions and, therefore the intercellular adhesion, promoting the barrier function of the gastrointestinal mucosa and preventing the passage of large molecules into the lamina propria (164-166). Probiotic bacteria may also regulate local and systemic immune responses through the production of IgA and anti-inflammatory cytokines; they must promote the antagonism against pathogenic bacteria and finally, they must enhance the synthesis of compounds with enzymatic

activity or metabolites beneficial to the host (165). Moreover, probiotic products' effectiveness is dose specific. In detail, based on the available scientific evidence, the minimum amount to obtain a temporary colonization of the intestine by a microbial strain is supposed to be at least 10^9 live cells per day (167). Therefore, the Italian Ministry of Health recommends that the portion of the product for daily consumption must contain 10^9 live cells for at least one of the strains presents.

The total load of live bacterial cells present in the probiotic products must be indicated on the label for each strain and must be guaranteed, together with the suggested storage methods, until the end of the shelf-life of the product. Commercial preparations containing probiotic bacteria may consist of one or more bacterial strains or genera and are available in capsules, freeze-dried powders, or contained in yogurt or fermented milk.

Due to their role in promoting health, probiotic products are widely present in individuals' daily diets (168). In this regard, it is necessary that the probiotic products currently on the market do not present formulation errors, such as insufficient levels of viable bacteria, lack of strains declared in the product and the presence of contaminating microorganisms. The qualitative analysis of probiotic products at company level traditionally takes place using conventional culture-dependent techniques, which are aimed at evaluating the ability of bacterial cells to grow and generate colonies in agar plates or liquid media. Through this approach, it is possible to discriminate only microorganisms that can be cultivated, but such techniques don't allow a direct taxonomic classification of the probiotic bacteria contained within the product. With this purpose, new methodologies have been proposed to verify probiotic formulations' integrity and quality in terms of taxonomic composition and cell viability (161, 169). Finally, analysis of the different ecosystems present in the human body, together with the genetic investigations of microorganisms, have made available new probiotic bacteria with specific characteristics and well-defined

mechanisms of action (158). The identification of next-generation probiotics (NGPs) that are selected and characterized using next-generation sequencing techniques and bioinformatics tools opens new avenues in the above-mentioned context (170). Future NGPs use directed to prevent and counteract gut microbiota dysbiosis may include “classical” probiotic bacteria such as bifidobacteria and lactobacilli but also other microbial groups of the human core gut microbiota (171).

I. Impact of the diet on the gut microbiome composition

The composition of the microbiota is dynamic and can be modulated by the exposure to prebiotic compounds and probiotic bacteria, but also individualized according to the influence of diet, which can be considered as a source of nutrients and prebiotics for gut microbiota as well as a potential vehicle of microbes (172) (Fig. 5).

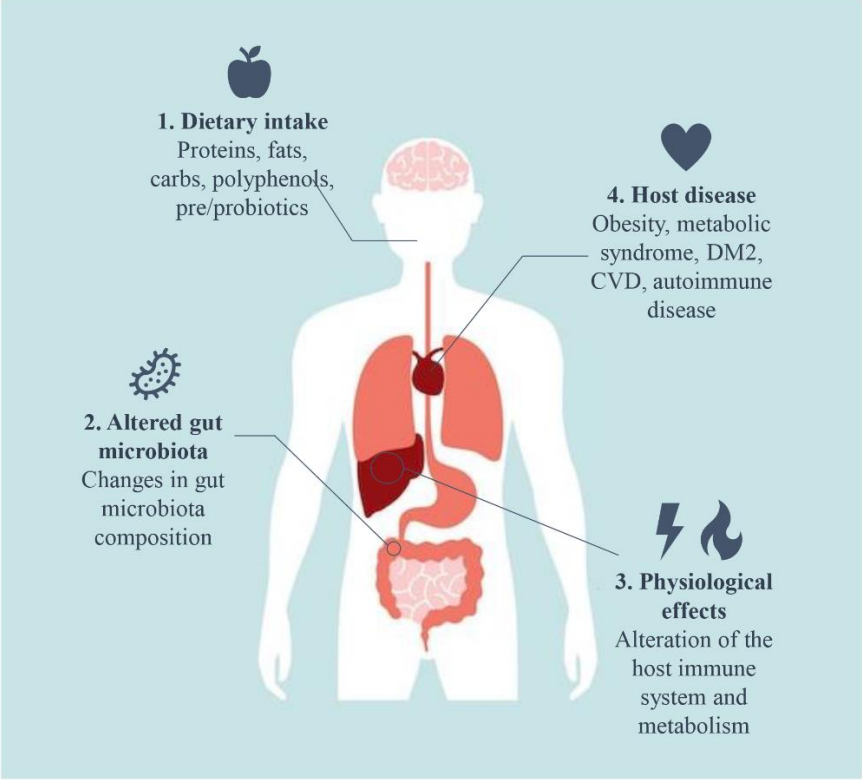


Figure 5: Impact of diet on human health and disease.

Thanks to metagenomic studies, it has been demonstrated that the composition of the human intestinal microbiota changes in various populations depending on different dietary habits (173).

Asian population, for example, displays a lower risk in the onset of IBD or even of tumors; this is because their diets contain antioxidant molecules, such as garlic, ginger and soy, that prevent cell proliferation and chronic inflammation in intestinal cells (174). On the other hand, in a Western-based diet, characterized by a low presence of fibers and a high occurrence of fat, fiber-degrading bacteria such as *Prevotella*, *Succinivibrio*, *Treponema* and bifidobacteria are generally slightly represented (98, 173). Conversely, a diet rich in whole grains not only exerts a bifidogenic effect by increasing bifidobacteria, but also increases the number of *Colinsella*, *Atopobium* and *Clostridium* (175, 176). Moreover, it has been demonstrated that an animal-based diet, composed of meats, eggs and cheeses, increased the abundance of bile-tolerant microorganisms such as *Alistipes*, *Bilophila* and *Bacteroides* and decreased the levels of those *Firmicutes* that metabolize dietary plant polysaccharides which are typically abundant in a plant-based diet, rich in grains, legumes, fruits, and vegetables (177). The most interesting data concern the influence of specific dietary components such as dietary fibers, which are glycans unable to be absorbed in the small intestine or digested by human enzymes (140, 141). In this context, dietary fibers promote the growth and the activity of butyrate-producing microorganisms such as *Roseburia* spp., *Eubacterium rectale* and *F. prausnitzii*. Furthermore, as already mentioned above, the presence of bifidobacteria and lactobacilli is stimulated by dietary fibers directly or indirectly by cross-feeding (178).

Little is known about the impact that proteins and fats have on the composition of the intestinal microbiota. Notably, it seems that fats can modulate the gut microbiota through bile acid composition and secretion (179). Recently, it has been demonstrated that most of the food introduced by the Western diet into the human body has limited benefits on the intestinal microbiota; this is also due to the excessive consumption of refined carbohydrates (180). Hence,

the growing demand not only for a more precise diet providing the essential nutrients to satisfy nutritional needs but also for a healthier one to avoid deficiencies.

Moreover, microorganisms residing in foods received much attention concerning their ability to colonize human gut microbiota modulating its composition and influencing human health (181, 182). In this context, the growing consumer interest in health has driven the development of functional foods, offering health benefits that extend beyond their nutritional value (183). Functional foods contain supplements or other additional ingredients like vitamins, minerals, fibers, or probiotic bacteria which have been shown to reduce inflammation and improve immune function and heart health (184). Fermented foods can be considered functional foods as well, due to their putative health benefits (184). These are functional foods containing microbes involved in the preservation of food through fermentation (185, 186), which is the process by which alcohols, carbon dioxide and/or organic acids are produced by microorganisms exploiting the sugars present in food material to produce energy (187). The accumulation of alcohol and organic acids, in addition to influencing the organoleptic qualities of the fermented food, plays a critical role in extending its shelf-life through the inhibition of the growth of undesirable spoilage and/or pathogenic microorganisms. Moreover, many fermented foods and beverages contain bacteria with potential health properties, added during the production process or originally inhabiting food microbiota, such as non-starter lactic acid bacteria (NSLAB) found in cheese, which can grow and proliferate in fermented foods (69). The ability of fermented foods to influence the gut microbiome has been widely demonstrated by several studies, which revealed strong associations between weight control and consumption of fermented dairy products (188), reduced risk of cardiovascular disease and type 2 diabetes in individuals consuming yogurt (189-192), and improved glucose metabolism associated with reduced muscle soreness following acute resistance exercise as a

consequence of consuming fermented milk (193). In addition, certain chemicals, such as polyphenols and dietary fibers present in fermented foods, can affect the host's gut microbiome directly, the latter of those leads to the production of SCFAs (194).

Therefore, the microbial transmission from food to the human intestine is well-known; recently, it has been demonstrated the existence of a potential horizontal transmission of bovine gut and milk bacteria to the human gut via the consumption of Parmesan cheese (195). In this context, it has been found that bacteria like *Bifidobacterium mongoliense* harbored by the bovine gut microbiota or by the housing environment modulate the microbiota of bovine milk and consequently that of cheese. Moreover, bacteria inhabiting Parmesan cheese may colonize and persist in the gut of those individuals who daily consume such product (195). However, while the modulation of the human gut microbiota driven by foods has been extensively studied, the microbiota of beverages, which are classified as foods, is still largely ignored.

Chapter 2

Outline of the thesis

The purpose of this Ph.D. thesis is to provide insights into the composition of different microbial communities through the application of metagenomic approaches to shed light on those factors affecting the host's gut microbiota. Gut microbiota is generally accepted as a forgotten organ of the human body and given its importance in human health, it needs to be properly studied. Indeed, given recent advances in the field of molecular biology, it has emerged that the gut microbiota plays a crucial role in determining the health status of the host throughout its entire life. Furthermore, alterations in the gut microbial balance have been associated with various diseases and disorders. In this regard, Next Generation Sequencing techniques have made it possible to understand the role played by host-microbe interactions on human health.

Chapter 3 investigates the performances of a DNA isolation saponin-based protocol commonly applied to different biological specimens collected from human donors to deplete host DNA which may disturb the assessment of the microbial population, specifically for diagnostic purposes. Shotgun metagenomic-based data showed that saponin allowed to remove most of the host's DNA in favor of bacterial DNA. However, further investigation revealed a drastically change in the detected microbial composition with an overall increase in the relative abundance of Gram-positive bacteria due to the reduction of Gram-negative, which appear to be more susceptible to saponin-induced lysis.

Chapter 4 illustrates a novel metagenomic pipeline, referred to as Probiotic Identity Card (PIC), able to reveal the bacterial composition of probiotic products and the genetic as well as genomic content of the strains included in the analyzed probiotic supplements. This pipeline involves whole metagenome shotgun analyses combined with flow cytometry assays, generating data which are useful to determine the integrity and the quality of the probiotic formulations.

Chapter 5 describes the microbiota harbored by 128 Italian PDO raw milk cheeses collected from different geographical locations and reveals that the microbial content of each cheese product is unique in taxonomical composition as well as metabolic and genetic properties, giving the final product specific organoleptic characteristics, probably correlated to the type of manufacturers' process and also environmental factors.

Chapter 6 discusses a still open question concerning the microbial composition of fresh potable water and its potential impact on human gut microbiota. In particular, a shotgun metagenomic analysis has been performed on water samples from public fountains and household taps of Parma city, allowing us to investigate their microbial biodiversity and to identify the presence of a conserved core tap water microbiota mainly represented by unknown bacterial species. Finally, genome reconstruction of the dominant bacterial genera of water microbiota allows tracing their presence into the human fecal microbiome, revealing a potential novel route of microbial transmission.

Chapter 3

Saponin treatment for eukaryotic DNA depletion alter the microbial DNA profiles by reducing the abundance of Gram-negative bacteria in metagenomics analyses

Longhi G, Argentini C, Fontana F, Tarracchini C, Mancabelli L, Lugli G.A, Alessandri G, Lahner E, Pivetta G, Turroni F, Ventura M, Milani C

The results of this chapter are under revision in Microbiome Research Reports.

Summary

Recent advances in microbiome sequencing techniques have provided new insights into the role/s of the microbiome on human health with potential diagnostic implications. However, these developments are often hampered by the presence of a large amount of human DNA interfering with the analysis of the bacterial content. Nowadays, there is a huge scientific literature focusing the interest on eukaryotic DNA depletion methods, which seem to be successful in removing host DNA in microbiome studies, even if a precise assessment of the impact on bacterial DNA is often missing.

Here, we have investigated a DNA isolation saponin-based protocol that is commonly applied to different biological matrices to deplete the released host DNA. The bacterial DNA obtained was used in order to assess the relative abundance of bacterial and human DNA, revealing that the inclusion of 2.5 % (wt/vol) saponin allowed the depletion of most of the host's DNA in favor of bacterial DNA enrichment. However, shotgun metagenomic sequencing showed inaccurate microbial profiles of the DNA samples, highlighting an erroneous increase of Gram-positive DNA. Even the application of 0.0125 % (wt/vol) saponin altered the bacterial profile by depleting Gram-negative bacteria resulting in an overall increase of Gram-positive bacterial DNA.

Originality-Significance Statement

The application of the saponin-based protocol drastically changes the detection of the microbial composition of human-related biological specimens. In this context, we revealed that saponin targets not only host cells but also specific bacterial cells, thus inducing a drastic reduction in the profiling of Gram-negative bacterial DNA.

Introduction

The microbiome research and especially the detection of microorganisms by molecular techniques has become a fundamental tool for the investigation of host-associated bacteria, such as those harbored by veterinary or human clinical samples (1, 2). Next Generation Sequencing (NGS) approaches now enable the identification of slow-growing, non-cultivable or non-viable bacteria contained in clinical specimens, without the need to rely on conventional culture-based identification methods based on the isolation of microorganisms by *in vitro* growth (3, 4). Initially, for DNA sequencing cost reasons, basic PCR amplicon systems were exploited providing the amplification of a specific gene for genus-level phylogenetic profiling purposes (5). The study of microbial composition is rapidly progressing due to recent advances in DNA sequencing platforms employed for shotgun metagenomics sequencing approaches, and concomitantly lowering of the costs of DNA sequencing, thus guaranteeing the investigation of species-level taxonomic profiles along with genetic and functional information of host-associated microbiomes (6).

However, despite the use of protocols specific for bacterial DNA extraction, the overwhelming amount of host DNA in many biological samples analyzed generate no reliable metagenomic data where the microbial sequences are hindered by the huge amount of eukaryotic DNA. High human-microbial DNA ratios have been reported for skin swabs as well as sputum (7), saliva (8), oral swab (9) vaginal samples (10) and human biopsies (11) making it difficult to investigate the resident microbial population. The ratio of human/microbial DNA can also be increased when samples belong to inflamed or infected sites due to influx of immune cells, tissue wounds, or necrosis (12).

Accordingly, the amount of bacterial DNA present in some biological samples can reach very low levels, such as on skin samples due to the cutaneous low pH and the continuous secretion of

antimicrobials (13) or biopsies where the bacterial content is mostly associated to tissue or mucosa (14). In this context, methods have been developed to enrich bacterial DNA, but they proved to be partially ineffective and compatible only with fresh specimens (15, 16). For this reason, the extraction of bacterial DNA from biological samples with a high contamination of host eukaryotic DNA is usually much more demanding in terms of target sequencing depth, and related costs, in order to compensate for the lower fraction of microbial DNA. Thus, optimization of the extraction and sequencing protocols are now mandatory (17).

Several approaches have been proposed in order to optimize DNA obtained from samples of different sources low in microbe's counts. Many commercially available microbiome-specific DNA extraction kits (18) and the use of a variety of additional steps employing detergents including saponin, Tween 20 and Triton X-100 (16), or Benzonase (19) are amongst the approaches that have been developed to deplete host DNA without apparently influencing the prokaryotic DNA. Nevertheless, while many studies have demonstrated that these approaches could reduce human DNA contaminations, few have pointed out the great impact it may have on the final microbial community structure that is delineated by shotgun metagenomics attempts. Among these approaches, different saponin percentages have been extensively tested and proposed as the golden standard for the depletion of eukaryotic DNA in highly contaminated samples with human DNA (20-22). However, an important limitation of the use of saponin is linked to the differential impact that this reagent produces on the DNA of the various bacteria and thus on the generation of artefacts in the relative abundance of the different microbial groups. In this regard, saponin has been previously reported to possess strong antimicrobial activity toward specific bacterial taxa both through *in vitro* and *in vivo* investigations (23).

For this reason, in this study we have carefully evaluated the saponin-based protocol for the extraction of DNA from different human sample specimens as a method to overcome high eukaryotic turnover and highlighted its main limits in terms of microbiome profile.

Experimental Procedures

Human samples collection. Regarding the biological specimens included in this study, for each human matrix was collected one individual sample. Specifically, the vaginal swab sample was gathered by a healthy woman. The oral sample consisted of a lingual swab, and sputum samples represent the thick mucus expelled from the lower airways (bronchi and lungs) through a deep cough. Saliva samples were the early morning sampling before teeth washing. Skin sample represented the forehead microbiota sampled with film dressing. Biopsies were different section of the stomach (gastric antrum and body) and the collection of the nasopharyngeal swab has been performed at Parma Day Hospital by the nursing staff. All the biological specimens included in this study were collected from different healthy donors and stored at -80° C until they were processed.

Signed informed consents were obtained from the individuals enrolled in this study.

Human DNA depletion and bacterial enrichment. Samples (1 ml) were centrifuged at 6,000*g* for 3 min, after which the supernatant was carefully removed, and cell pellets were incubated with Phosphate Buffer Saline (PBS) and saponin (Sigma-Aldrich, MO, USA) at the required concentration (2.5% wt/vol, 2% wt/vol, 1.5% wt/vol, 1% wt/vol, 0.5% wt/vol, 0.1% wt/vol, 0.05% wt/vol, 0.025% wt/vol and 0.0125% wt/vol) at room temperature for 10 min. Following this incubation, 350 µl of sterile water was added and incubation was continued at room temperature for 30 s, after which 12 µl of 5 M NaCl was added to deliver an osmotic shock, lysing the damaged host cells. Samples were next centrifuged at 6,000*g* for 5 min, with the supernatant removed and the pellet resuspended in 100 µl of PBS. Turbo DNase buffer (Thermo Fisher Scientific, USA)

with Turbo DNase enzyme (Thermo Fisher Scientific, USA) were added at 37°C for 30 min to promote host cell lysis. Finally, the host DNA depleted samples were washed two times with decreasing volumes of PBS (1 ml and 800 µl). After the final wash, the samples were centrifuged at 6,000g for 3 min, the supernatant discarded, and the pellet resuspended in 600 µl of PBS. We included three cycles of bead-beating followed by 3 steps on ice.

DNA extraction and quantification. The DNA was extracted from the specimens using commercially available kits following the manufacturer's instructions. For bacterial DNA extraction, the best performing kits were employed on the basis of the biological matrix of origin. QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) was exploited to extract bacterial DNA from sputum, saliva, oral and nasopharyngeal swabs and biopsies and ZymoBIOMICS DNA Miniprep Kit (Zymo Research, D4300) for vaginal swabs. Then, the DNA concentration and purity were investigated employing a Picodrop microtiter Spectrophotometer (Picodrop, Hinxton, UK).

Quantitative Real-Time PCR for bacteria and human DNA. With the aim to evaluate the bacterial and eukaryotic abundance in samples treated or not with saponin and DNase we performed a qPCR approach with specific primers for bacteria and human, i.e., Probio_uni (5′-CCTACGGGRSGCAGCAG-3′) and Probio_rev (5′-ATTACCGCGGCTGCT-3′) for bacterial 16S rRNA (24), bGLB+ (5'-ACACAACTGTGTTCACTAGC-3') and bGLB- (5'-CAACTTCATCCACGTTCACC-3') for the human β-globin (25). DNA from the different biological samples was extracted and diluted to a concentration of 10 ng/µl. qPCR was performed using PowerUp SYBR Green Master Mix (ThermoFisher Scientific, US) on a CFX96 system

(BioRad, CA, USA) following previously described protocols (26). PCR products were detected with SYBR green fluorescent dye and amplified according to the following protocol: one cycle of 50°C for 2 min, followed by one cycle of 95°C for 2 min, followed by 40 cycles of 95 °C for 15 s, 55-60°C for 15 s and 72°C for 1 min. The melting curve was 65 °C to 95 °C with increments of 0.5 °C/s. In each run, negative controls (no DNA) were included. A standard curve was generated using the CFX96 software (BioRad) and chromosomal DNA belonging to untreated biological matrices and *Bifidobacterium longum* were used as qPCR standards for the eukaryotic and bacterial detection, respectively.

Shallow shotgun sequencing. According to the manufacturer's instructions, DNA library preparation was performed using the Nextera XT DNA sample preparation kit (Illumina, San Diego, CA, USA). First, one ng input DNA from each sample was used for the library preparation, which underwent fragmentation, adapter ligation, and amplification. Then, Illumina libraries were pooled equimolarly, denatured, and diluted to a concentration of 1.5 pM. Next, DNA sequencing was performed on a MiSeq instrument (Illumina) using a 2X 250 bp Output sequencing Kit together with a deliberate spike-in of 1% PhiX control library.

Short read taxonomic classification. Sequenced paired-end reads of each sample were subjected to a filtering step removing low-quality reads (minimum mean quality score 20, window size 5, quality threshold 25, and minimum length 100) using the fastq-mcf script (https://github.com/ExpressionAnalysis/ea-utils/blob/wiki/FastqMcf.md) to analyze high-quality sequenced data only. Then, employing the BWA aligner, an additional filtering step was performed

to remove the reads mapping against the *Homo sapiens* genome sequence, thus removing possible contaminating human DNA (27). Retained reads were taxonomically classified through the METAnnotatorX2 pipeline (28), using a set of databases of reference genomes whose taxonomy was previously validated to maximize the accuracy of homology-based taxonomic classification of reads (29).

Data availability. Shotgun metagenomics data are accessible through SRA study accession number PRJNA864589.

Results and Discussion

Evaluation of bacterial and human abundance in biological samples with high content of eukaryotic DNA. In order to preliminary estimate the effectiveness of the best approach for the bacterial DNA extraction aimed at reducing the carryover of human host DNA, we tested different amounts of saponin (2.5 % wt/vol, 1 % wt/vol, 0.1 % wt/vol and 0.0125 % wt/vol) on a saliva sample collected from a healthy human donor. Saliva was chosen as an optimal test case due to the consistently high percentage of human DNA (~ 90%) that is determined by shotgun metagenomic sequencing (8). Treatment with saponin aimed to induce lysis of eukaryotic cells and it was followed by adding DNase to the DNA extract in order to remove eukaryotic DNA prior to bacterial DNA extraction with a commercial kit after mechanical bead-beating lysis. Specifically, to assess the bacteria-human DNA ratio, qPCR was performed on the total DNA obtained using specific PCR primers targeting the bacterial 16S rRNA gene, while human DNA level present into the saliva sample was determined by absolute quantification of the β-globulin gene (25). In addition, this analysis was performed on DNA extracted from the same saliva sample that did not undergo the depletion process, as a reference control, to rule out depletion as a potential cause of missed/additional bacterial detection.

Across all saponin amounts used, 2.5 % (wt/vol) of detergent was found out to successfully deplete most of the human DNA (Figure 1a). Indeed, in a saliva sample treated with 2.5 % (wt/vol) of saponin, the human DNA decreased to a concentration of 8.01 gene copy number/ml compared to the untreated sample in which the host's DNA was around $7.13*10^1$ gene copy number/ml (Figure 1a). Conversely, host DNA abundance in the same saliva samples treated with 1 % (wt/vol), 0.1 % (wt/vol) and 0.025 % (wt/vol) of saponin remained relatively stable at a concentration of 2-3 * 10^1 gene copy number/ml (Figure 1a).

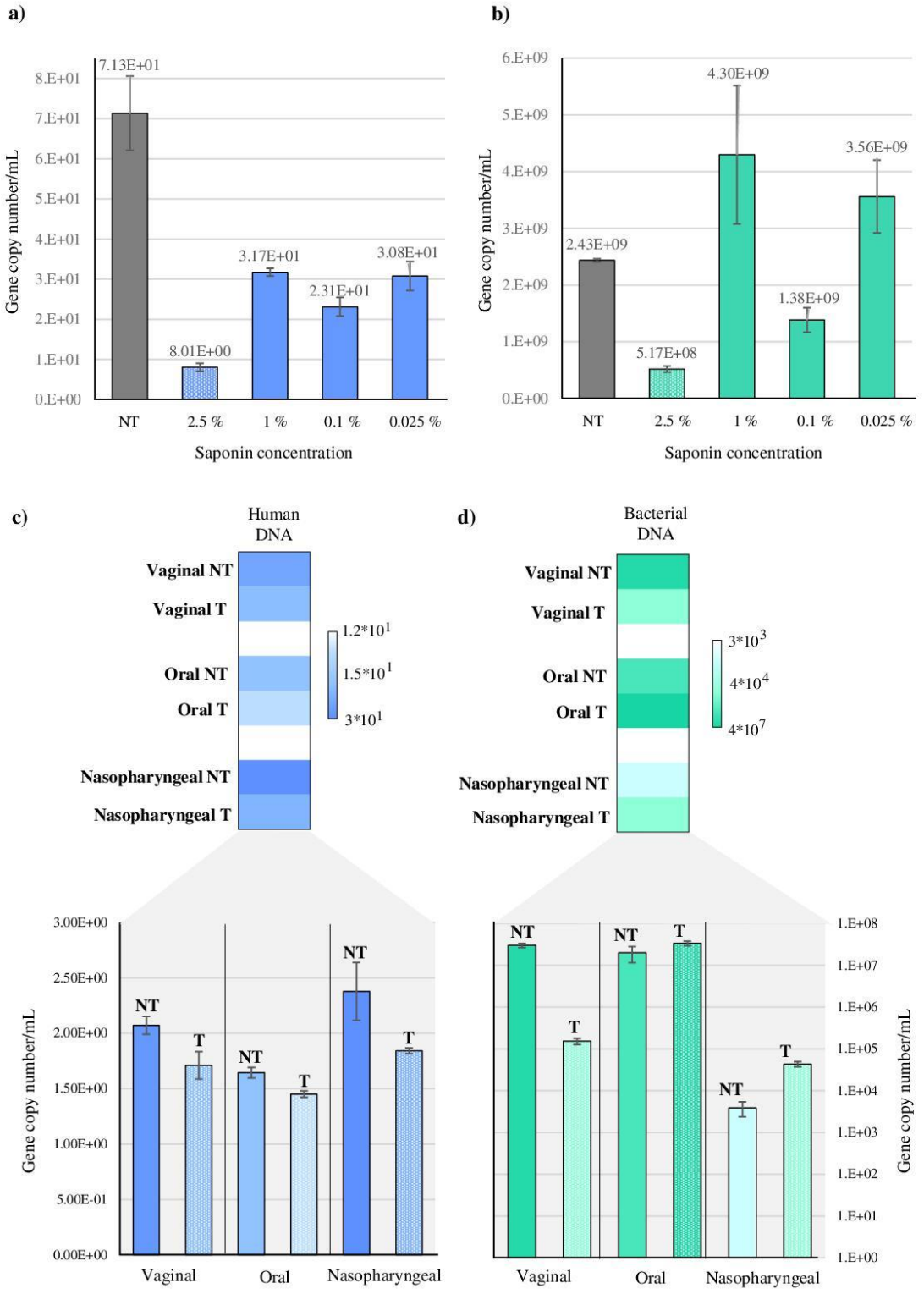

a)
Gene copy number/mL
8.E+01
7.E+01
6.E+01
5.E+01
4.E+01
3.E+01
2.E+01
1.E+01
0.E+00
7.13E+01
8.01E+00
3.17E+01
2.31E+01
3.08E+01
NT
2.5 %
1 %
0.1 %
0.025 %
Saponin concentration
b)
Gene copy number/mL
6.E+09
5.E+09
4.E+09
3.E+09
2.E+09
1.E+09
0.E+00
2.43E+09
5.17E+08
4.30E+09
1.38E+09
3.56E+09
NT
2.5 %
1 %
0.1 %
0.025 %
Saponin concentration
c)
Human DNA
Vaginal NT
Vaginal T
Oral NT
Oral T
Nasopharyngeal NT
Nasopharyngeal T
1.2*10^1
1.5*10^1
3*10^1
Gene copy number/mL
3.00E+00
2.50E+00
2.00E+00
1.50E+00
1.00E+00
5.00E-01
0.00E+00
NT
T
Vaginal
Oral
Nasopharyngeal
d)
Bacterial DNA
Vaginal NT
Vaginal T
Oral NT
Oral T
Nasopharyngeal NT
Nasopharyngeal T
3*10^3
4*10^4
4*10^7
Gene copy number/mL
1.E+08
1.E+07
1.E+06
1.E+05
1.E+04
1.E+03
1.E+02
1.E+01
1.E+00
NT
T
Vaginal
Oral
Nasopharyngeal

Figure 1

Figure 1. Abundance of bacterial and human DNA in different biological samples with high content of eukaryotic DNA evaluated through qPCR. Panel (**a**) shows absolute quantification of host DNA (light blue) and panel (**b**) indicates the corresponding absolute quantification of bacterial DNA (green) in saliva samples treated with different amounts of saponin compared to the untreated counterpart (dark gray). Each pillar represents the average quantity of human and bacterial DNA content ± standard deviation. Panel (**c**) displays the absolute quantification of host DNA (light blue) in a vaginal, oral and nasopharyngeal swab treated with 2.5 % wt/vol concentration of saponin and subsequently treated with DNase prior to proceed with microbial DNA extraction using the best performing commercial kit. Panel (**d**) shows the bacterial DNA (green) counterpart in vagina, oral, and nasopharyngeal samples. Every pillar represents a biological sample treated with saponin (T) or its untreated counterpart (NT). The color scales (blue and green) at the top of panels (**c**) and (**d**) indicate the increase in the amount of human or bacterial DNA in the analyzed samples.

Furthermore, the 2.5 % (wt/vol) saponin treated sample showed a $5.17*10^8$ gene copy number/ml bacterial content compared to the untreated saliva in which bacterial DNA had a concentration of $2.43*10^9$ gene copy number/ml (Figure 1b). These findings showed that saponin is able to lysate human cells by reducing the eukaryotic DNA content within the sample, but high amounts of such detergent (e.g., 2.5 % wt/vol) also acted on bacterial cells, decreasing in part their abundance. Given the impact of saponin on the bacterial content, we suggest to preliminarily test the biological matrix according to the purpose of the study in order to correct the quantity of saponin needed to be used. Specifically, treatment with saponin is risky if applied to biological samples originally low in bacterial content, because the latter could be totally lost. This depletion protocol, on the other hand, is very useful if applied to samples with a high content of bacterial, in order to reduce the contamination of host-derived DNA.

Investigation of how the saponin-based protocol impact on different biological matrices.

Preliminary results were validated on other biological specimens from different human body sites rich in eukaryotic DNA such as from vagina, oral cavity, and nasopharynx. Each sample was processed with a protocol reported by Charalampous *et al.* on Nature Biotechnology in 2019 employing a 2.5 % wt/vol concentration of saponin for eukaryotic DNA removal (20) and subsequently the DNA extracts were treated with DNase prior to proceed with standard microbial DNA extraction. For the bacterial DNA extraction from different biological matrices, distinct commercially available kits were used as described in the Experimental Procedures section. The amount of bacterial and eukaryotic DNA present in the nucleic acid extracts was evaluated by qPCR using specific primers for the discrimination of bacterial 16S rRNA and β-globulin genes.

Data collected revealed that saponin treatment induced a depletion of eukaryotic DNA in all the samples analyzed. (Figure 1c). Notably, after the saponin treatment, host DNA content decreased by 17.5 % in vaginal, 11.7 % in oral and 22.6 % in nasopharyngeal swab samples in respect to the same untreated samples (Figure 1c). In contrast, bacterial content seemed to slightly increase from 10^3 to 10^4 gene copy number/ml in the nasopharyngeal swab as well as in the oral and in the nasopharyngeal samples after saponin-based depletion, suggesting that lower host's DNA abundance may support the recovery of a higher amount of bacterial DNA in some specific biological matrices in contrast to what observed for saliva (Figure 1d). However, confirming the previous findings, saponin depletion protocol seemed to impact also on bacterial cells promoting their lysis in biological samples such as vaginal swabs. In this case, probably due to the large amount of bacterial DNA, bacterial content decreased from 10^7 to 10^5 gene copy number/ml in the saponin-treated vaginal swab (Figure 1d). Remarkably, the results obtained confirmed what observed for the saliva samples, i.e., saponin targets mainly human DNA, showing great performances in the removal of eukaryotic DNA from the biological samples, but also sometimes impacts on the bacterial DNA content. As we mentioned previously, the observed differences about bacterial population reduction can be considered as biological matrix dependent. Thus, we suggest to preliminary test each biological matrix included in a study to identify the best depletion approach in terms of saponin concentration according to the study purpose and the bacterial content of the specimens analyzed, before applying the saponin treatment to all samples.

Validation of the saponin DNA extraction protocol by shallow shotgun metagenomics of a wider range of biological matrices. The shotgun metagenomics approach allows to randomly sequence all the DNA fragments obtained from DNA extraction protocols, thus allowing to

quantify the prokaryotic/eukaryotic DNA ratio through specific bioinformatic applications (29). To validate the outcome of our qPCR analyses, shallow shotgun metagenomic sequencing (30) was performed, a specific low depth shotgun metagenomics technique that provides a reliable bacterial profile at species level (29). DNA samples from different biological matrices rich in eukaryotic DNA, processed with the extraction protocol involving 2.5 % (wt/vol) saponin and DNase followed by bead-beating and kit extraction steps, was submitted to DNA sequencing. In order to provide a more comprehensive overview of the human-associated microbiota, we included samples rich in eukaryotic DNA such as vaginal swab, saliva, skin, sputum and nasopharyngeal swab samples (Table 1).

Table 1. Filtering table of the analyzed DNA samples from different biological matrices rich in eukaryotic DNA, processed with the extraction protocol involving 2.5% (wt/vol) saponin and DNase followed by bead-beating and kit extraction steps. Finally, the DNA obtained was submitted to shallow shotgun sequencing.

Biological samples	Sequenced reads produced	High quality reads	Reads retained after *Homo sapiens* filtering	% Filtered *Homo*
Vaginal-NT	154,701	151,193	4,266	97.18%
Vaginal-T	222,202	211,864	159,017	24.94%
Gastric antrum-NT	101,522	99,529	494	99.50%
Gastric antrum-T	59,749	56,936	251	99.56%
Gastric body-NT	136,117	133,250	1,214	99.09%
Gastric body-T	143,220	139,737	199	99.86%
Saliva-NT	134,011	131,338	30,156	77.04%
Saliva-T	225,114	218,562	208,893	4.42%
Skin-NT	242,037	233,438	128,078	45.13%
Skin-T	236,997	227,914	163,985	28.05%
Sputum-NT	77,837	52,182	9,776	81.27%
Sputum-T	208,337	203,745	200,275	1.70%
Nasopharingeal-Swab-NT	78,150	71,634	3,613	94.96%
Nasopharingeal-Swab-T	4,184	4,017	3,933	2.09%

Sequencing output constituted of an average of circa 144,585 pair-end reads per sample (Table 1), thus allowing accurate assessment of the microbial profile inhabiting each specimen (30). This taxonomic survey better illustrate what has already been shown in the qPCR performed. Each biological matrix treated according to the depletion approach was compared to its untreated counterpart, with the aim of investigating the impact of the saponin method on the bacterial profile of each sample (Table 1).

The percentage of *Homo sapiens* reads obtained after the *in silico* filtering step in treated vaginal swab, saliva, skin, sputum and nasopharyngeal swab samples ranged from 1% to a maximum of 28.05 % (Table 1), with a reduction ranging from 45.1 % to 97.2 % respect to untreated samples. Thus, indicating that the saponin-based protocol was allowing a successful depletion of host DNA.

Investigating the microbial profiles of the saponin achieved DNA samples. Shallow shotgun metagenomic sequencing was also exploited to identify the microorganisms that populate biological samples at species level and to validate their microbial taxonomic composition. Taxonomic profiles obtained through METAnnotatorX2 software demonstrated that the skin, vaginal and nasopharyngeal swab samples processed with 2.5 % (wt/vol) saponin had the most similar bacterial pattern to their untreated counterparts. To evaluate the divergence of treated and untreated profiles, we employed a taxonomic variation index (TVI), consisting in the absolute sum of positive and negative relative abundance differences observed for each microbial taxon, thus ranging from 0 for identical profiles to 200 for completely different taxonomic profiles. For the vaginal, skin, nasopharyngeal swab and saliva sample, the retrieved the TVI was 29.1, 5.05, 23.6 and 1.74 respectively (Table S1). Specifically, this taxonomic survey revealed that, amongst the most impacted taxa, *Bifidobacterium scardovii* was present in the untreated vaginal sample at a relative abundance of 49.8 %, compared to the 64.3 % in the treated counterpart (Table S2) (Figure S1). In contrast, the relative abundance of *Bifidobacterium* spp., *Lactobacillus acidophilus* and *Lactobacillus gasseri* in vaginal swab remained mostly constant even after the depletion protocol (Table S2) (Figure S1). Accordingly, *Cutibacterium* spp. in the skin sample and *Corynebacterium* spp. in nasopharyngeal swabs were present at the same relative abundance in both saponin-treated and saponin-untreated sample (Table S2) (Figure S1). In these biological samples, only a few

bacterial species (~5 depending on specimen) present at low abundance (lower than 2 %) in the control were absent in the saponin-treated sample (Table S2). Remarkably, this observation suggested that saponin treatment induce different effect based on the microbial species present in the original biological sample.

Nevertheless, the same limited impact on the bacterial DNA content was not observed for other biological matrices such as biopsies (gastric antrum and body), saliva and sputum samples for which the same saponin-based extraction method induced a substantial alteration of the whole microbial taxonomic profile. In this case, the taxonomic variation index between treated samples and controls was 190.3 for the gastric antrum and 180.5 for the gastric body biopsies, 129.7 for the sputum respectively (Table S1). Consistently, 14 bacterial taxa with a prevalence between 0.5 % to 25.7 % were lost in the saponin-treated biopsy of the gastric antrum, 33 in the biopsy of the gastric body, 17 in the saliva and 12 in the treated sputum (Table S2) (Figure S2). Although with a relative abundance greater than 2 %, the most represented bacterial taxa in the saponin-treated samples range from four bacterial taxa in the biopsies to 15 and 11 in the saliva and sputum respectively (Table S2) (Figure S2). In saliva, sputum and biopsy specimens not treated with saponin, several members of *Haemophilus*, *Neisseria*, *Prevotella* and *Veilonella* genus were present with a relative abundance higher than 10 % and were the most representative bacterial taxa of these biological samples, while in the respective samples treated with saponin they fell under the limit of detection (Table S2) (Figure S2).

In contrast, a number of bacterial taxa not detected in the untreated specimens appeared in the profile of saponin-treated samples, with few cases (such as *Streptococcus* spp. and *Actinomyces* spp.) where these taxa even became the dominant ones in the retrieved taxonomic profiles (Figure

S2) (Table S2). Altogether, these data highlighted the impact of the saponin-based eukaryotic DNA depletion method, which can markedly alter the retained bacterial DNA profile.

In this context, we noticed that samples with a substantially high bacterial load were less compromised by the depletion protocol. Conversely, when processing samples poor in bacterial DNA such as mucosa, saliva and sputum, the saponin treatment risk to further reduce the already low bacterial population, resulting in the sequencing of a lower number of bacterial DNA sequencing which causes the loss of the minor component of the bacterial population in the taxonomic profiles retrieved after data analysis. Therefore, these results have greater impact on studies in which the purpose is represented by the comparison and the analysis of complex bacterial populations constituted by a high number of microbial species at low relative abundance. For this reason, preliminary test on the biological matrices included in each study should be performed.

On the other hand, it should be considered that studies focused specifically on the detection of one or few targets microbial taxa in simple microbial communities are less affected by the use of saponin depletion protocol, which may represent an extremely useful tool by increasing the microbial/host DNA ratio. For example, in case of pathogen detection for clinical diagnosis, the use of saponin would be highly recommended to remove the large amount of human DNA present in the biological samples in particular when applying shotgun or PCR/amplicon DNA sequencing approaches targeting and amplifying specific microbial taxa.

Saponin/DNase treatment affects the Gram-positive/Gram-negative ratio of the retrieved metagenomic profiles. Microbial profiling revealed a different response to the saponin treatment of Gram-positive and Gram-negative bacteria in the biological samples analyzed. More

specifically, saponin seemed to have an impact on the relative abundance of Gram-negative bacteria that prompt a concomitant increase in the proportion of Gram-positive microbes' respect to the whole population. It has been generally accepted that bacteria differ in their susceptibility to cell lysis based on the molecular structures composing their superficial biological structures (31). In contrast to Gram-positive, Gram-negative bacteria in addition to the cellular membrane possess a much thinner cell wall covered by an external membrane. This has been reported to increase susceptibility to specific cell lysis methods, thus explaining the drastic decrease of the number of Gram-negatives' DNA after saponin/Dnase treatment (32, 33).

Supporting this hypothesis, we observed that the sum of the relative abundance of Gram-negative bacteria in the antrum and in the gastric body biopsies was totally depleted after saponin treatment from 91.6 % and 79.7 % respectively to 0% in both. Furthermore, in saliva and sputum samples where Gram-negative bacteria represented 74.1 % and 48.8 %, respectively of the total amount of bacterial content in the control samples, their presence was completely depleted after treatment with saponin, due to a dominance of Gram-positive bacteria (Figure 2a) (Table S3). On the contrary, the biological samples where the microbial populations were mainly represented by Gram-negative bacteria, such as skin and nasopharyngeal swab, were the less affected due to a reduced shift in the Gram-positive/Gram-negative ratio, thus explaining the limited impact of saponin treatment on the taxonomic profiles obtained for these biological matrices (Figure 2a) (Table S3).

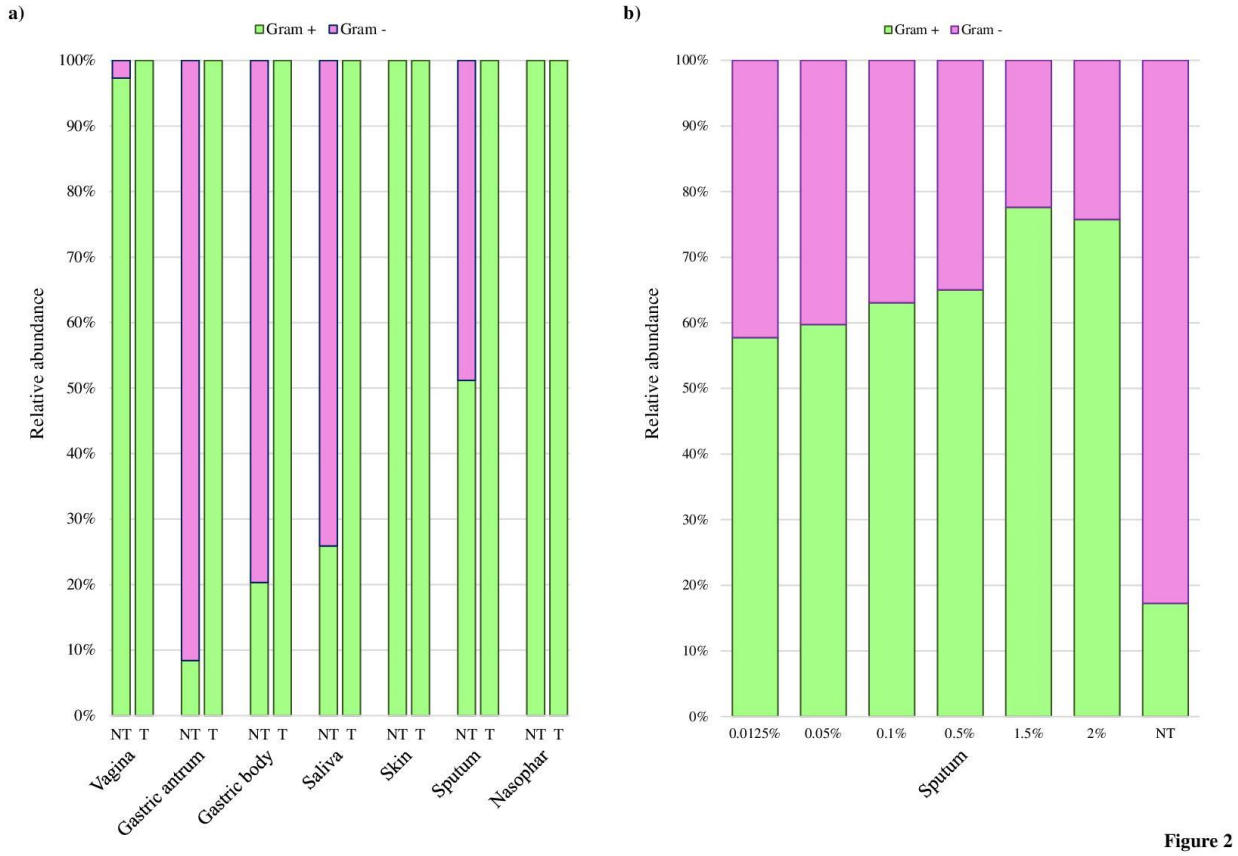


Figure 2. Gram-negative and Gram-positive bacterial ratio in different biological samples before and after saponin-based depletion protocol. Panel (**a**) indicates the percentages of Gram-negative (pink) and Gram-positive bacteria (green) on the total amount of bacterial DNA profiled in each biological sample before (NT) and after 2.5 % (wt/vol) saponin-based depletion (T). Panel (**b**) represents the percentages of Gram-negative (pink) and Gram-positive bacteria (green) detected in a sputum sample treated with different saponin amounts.

Overall, these data underline that saponin is able to markedly lower the Gram-positive/Gram-negative ratio in the recovered bacterial DNA, thus altering the post-sequencing microbial profiles.

In this regard, a preliminary taxonomic investigation of the microbial population is fundamental to evaluate if the use of saponin will be acceptable in terms of alteration of the taxonomic profiles.

In fact, in the case of samples with an overall Gram-positive bacterial population, alterations by saponin treatment will be more contained and acceptable in terms of accuracy of data analysis.

Comparison of the effect of different saponin percentages on the microbial profile. Although 2.5 % (wt/vol) is the generally used amount of saponin in the large part of the protocol for eukaryotic DNA removal, we decided to evaluate whether lower amount of this detergent could avoid the bias in the estimation of Gram-negative/Gram-positive bacteria ratio in the biological samples assayed. With this purpose, DNA isolated from the same sputum sample was processed with variable concentrations of saponin, i.e., 0.0125 % (wt/vol), 0.05 % (wt/vol), 0.1 % (wt/vol), 0.5 % (wt/vol), 1.5 % (wt/vol) and 2 % (wt/vol). Successively, DNA obtained from microbial DNA extraction was subjected to shallow shotgun metagenomic analysis.

Notably, the analysis of the achieved metagenomic data revealed that saponin had an impact on both human DNA abundance and Gram-positive/Gram-negative ratio directly proportional to the amount used. In fact, host DNA content ranged from 36.4 % in the 0.0125 % to 3.2 % in the 2 % (wt/vol) saponin treated saliva, compared to the control sample in which the host DNA content was 53.2 % (Table 2).

Table 2. Filtering table of the analyzed sputum samples processed with the extraction protocol involving different saponin amount, DNase and bead-beating treatments, with the aim of testing the effect of lower detergent concentrations on the estimation of the bias in the ratio of Gram-negative/Gram-positive bacteria.

7 Sample	Sequenced reads produced	High quality reads	Reads retained after *Homo sapiens* filtering	% filtered *Homo*
Sputum-0.0125%	51453	48877	31094	36,38%
Sputum-0.05%	34712	33795	31860	5,73%
Sputum-0.1%	41873	40807	39754	2,58%
Sputum-0.5%	13420	13056	12774	2,16%
Sputum-1.5%	34688	33633	32716	2,73%
Sputum-2%	37528	36418	35249	3,21%
Sputum-NT	38545	37556	17595	53,15%

To better detail the impact of saponin concentration, we focused our interest on the key Gram-negative bacterial taxa previously identified as dominant in the untreated sputum sample. In detail, the Gram-negative *Prevotella histicola* was present in the control sample with a relative abundance of 19 % and decreased from 9.1 % to 5 % respectively in the sample treated with 0.0125 % (wt/vol) and 2 % (wt/vol) of saponin (Table S4) (Figure S3). Similarly, the Gram-negative *Veillonella atypica* decreased from an initial relative abundance of 8.1 % in the untreated to 4.5 % and 2.6 % respectively in the 0.0125 % (wt/vol) and 2 % (wt/vol) treated samples (Table S4) (Figure S3). In contrast, the Gram-positive *Streptococcus salivarius* was detected as a dominant bacterial taxon in the sputum treated with 2 % (wt/vol) of saponin, but its relative abundance progressively decreased from 8.8 % in the DNA extracted with 0.0125 % (wt/vol) of saponin to 1.8 % in the control sample,

thus suggesting that its increase in relative abundance after saponin treatment was imputable to a reduction of the Gram-negative population (Table S4) (Figure S3). Similarly, the Gram-positive *Schalia* (unknown species) was present in the 2 % (wt/vol) treated saponin sample with a relative abundance of 8 % and decreased to 5.2 % due to 0.05 % (wt/vol) treated saponin respect to an abundance of 2.4 % in the control. Moreover, several other members of the *Prevotella* and *Veillonella* genera, as well as unknown species of the *Lancefieldella* and *Schalia* genera decreased with the increasing amount of saponin, while *Actinomyces* unknown species and other members of the *Streptococcus* genus increased, reinforcing the findings noticed above regarding the impact of the depletion protocol on the taxonomic profile (Table S4) (Figure S3).

These latest results confirmed that saponin treatment have an impact on Gram-positive/Gram-negative ratio. Gram-negative percentages decreased with the increasing of the saponin amount used, thus inducing a concomitant increase in the relative abundance of Gram-positive bacteria in the overall taxonomic profile. Indeed, Gram-negative ranged from an average of 42.3 % in the 0.0125 % (wt/vol) saponin-treated samples, to an average of 35 % and 24.3 % in the 0.5 % (wt/vol) and 2 % (wt/vol) saponin treated samples, respectively, compared to the starting 82.3 % in the untreated saliva sample (Table S5) (Figure 2b).

Remarkably, despite these data revealed that even low saponin amounts are able to modulate the retrieved bacterial DNA in terms of Gram-negative taxonomic representation, saponin could be efficiently employed for the depletion of human DNA when targeting the detection of microbial populations dominated by Gram-positive bacteria or when targeting specific microbial taxa such as pathogens in clinical contexts. Thus, data retrieved in this study remarked the need for preliminary tests on the biological matrices under investigation in order to assess the feasibility of the saponin protocol use.

Conclusions

In the framework of this study, we investigated the performances of the saponin-based eukaryotic DNA depletion approach on different biological matrices and the impact on the correlated microbial taxonomic profiles.

Overall, the host DNA depletion method tested successfully reduced the amount of human DNA in the nucleic acid extracts while not or barely reducing the bacterial DNA content. Nevertheless, we reported that the saponin-based protocol drastically changed the detected microbial composition harbored by the specimens analyzed. In this context, we revealed that saponin targets not only host cells but also specific bacterial cells, thus inducing a drastic reduction of Gram-negative bacteria DNA, probably due to the interaction of this detergent with the external membrane of the bacterial cell.

It is important to highlight that our study focuses on a single approach to deplete eukaryotic DNA from biological samples known to encompass large amount of host DNA, and it displays important limitations. However, even if the saponin based approach is one of the most used methods, there are also other protocols that are pursuing the same aim (18, 19, 34). Even in these cases, it will be necessary a very careful evaluation of their potential limits in order to avoid alteration of the retrieved microbial DNA. In fact, this represents a very serious issue that might affect the reliability of the microbial profiles and the biological meaning of the associated metagenomic studies that are using these specific DNA extraction protocols.

Based on the results obtained in our study, we highlighted that the use of saponin should be avoided if specific evaluation of the microbial/host DNA ratio and complexity of the targeted microbial

population are not known. In fact, exploring complex microbial communities with low bacterial load or communities composed by mixed Gram-positive and Gram-negative bacteria may result in issues in accuracy of the retrieved results. For this reason, we suggest that for each study intending to apply saponin treatment, it is necessary to introduce detailed preliminary tests on the biological matrices investigated in order to assess feasibility and utility of the saponin host DNA depletion approach.

In this context, a further novel approach that can be employed to overcome contaminations derived from high amounts of host DNA as an alternative to saponin could be represented by the SQK-RPB004 kit combined with the SQK-LSK109 kit proposed by Oxford Nanopore Technologies consisting in the introduction of a specific enrichment or depletion of target DNA directly during sequencing by simply providing reference fasta sequence (35). Nevertheless, it should be considered that this approach may vastly impact on the amount of sequencing data retrieved.

Supplementary figures

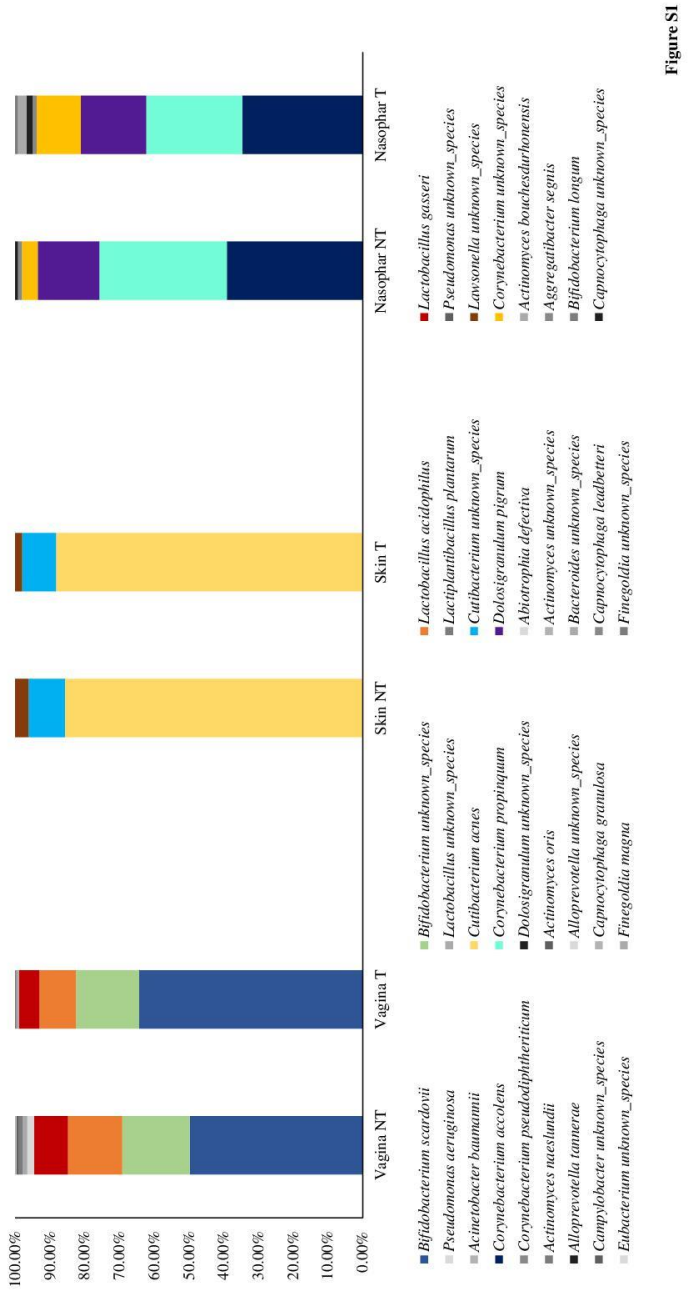


Figure S1. Taxonomic microbial profiling of biological samples showing limited biases in microbial taxonomy after 2.5 % (wt/vol) saponin treatment. The histogram displays the relative abundance of each microbial species identified in the vagina, skin, and nasopharynx samples. Both saponin-treated (T) and saponin-untreated samples (NT) are reported for each matrix. From left to right, the color-coding order of the legend reflects average abundances from largest to smallest.

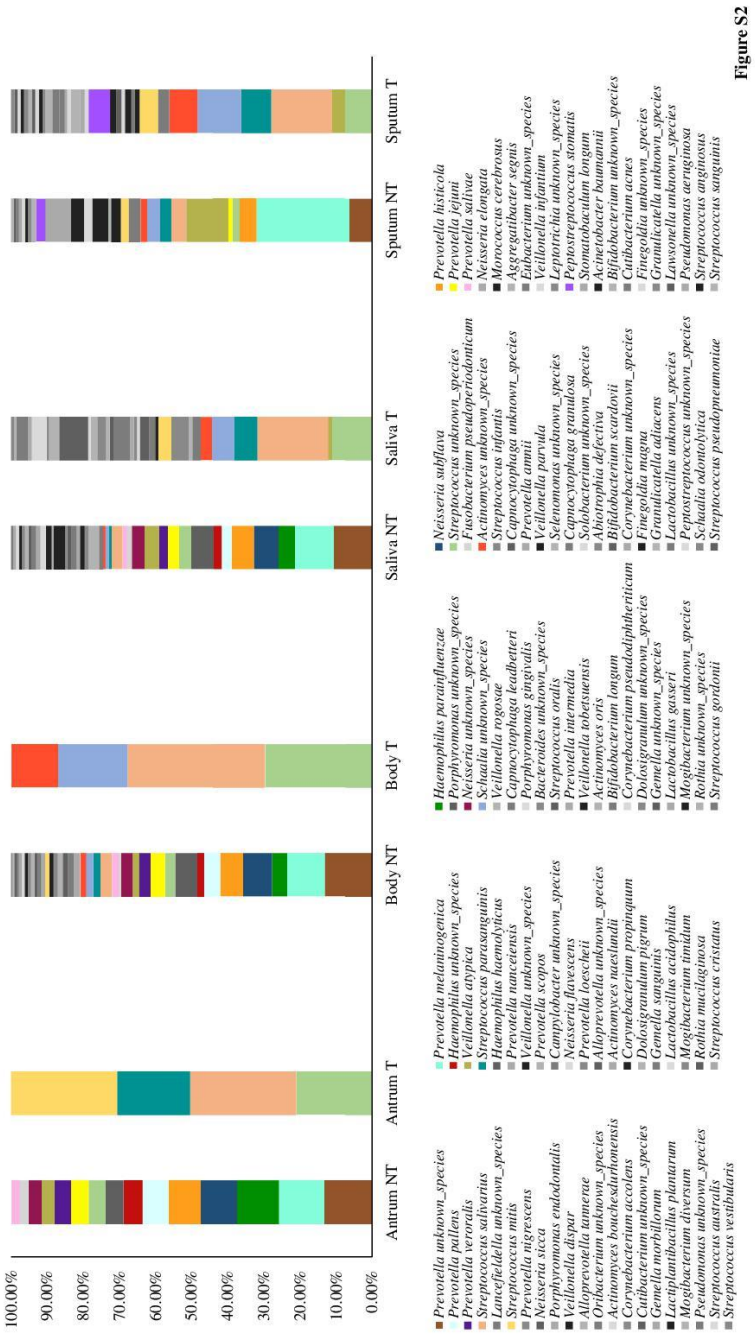


Figure S2. Taxonomic microbial profiling of biological samples showing marked biases in microbial taxonomy after 2.5 % (wt/vol) saponin treatment. The histogram displays the relative abundance of each microbial species identified in the biopsies, gastric antrum and body, saliva and sputum. Both saponin-treated (T) and saponin-untreated samples (NT) are reported for each matrix. From left to right, the color-coding order of the legend reflects average abundances from largest to smallest. Most representative bacterial species are indicated with different colors.

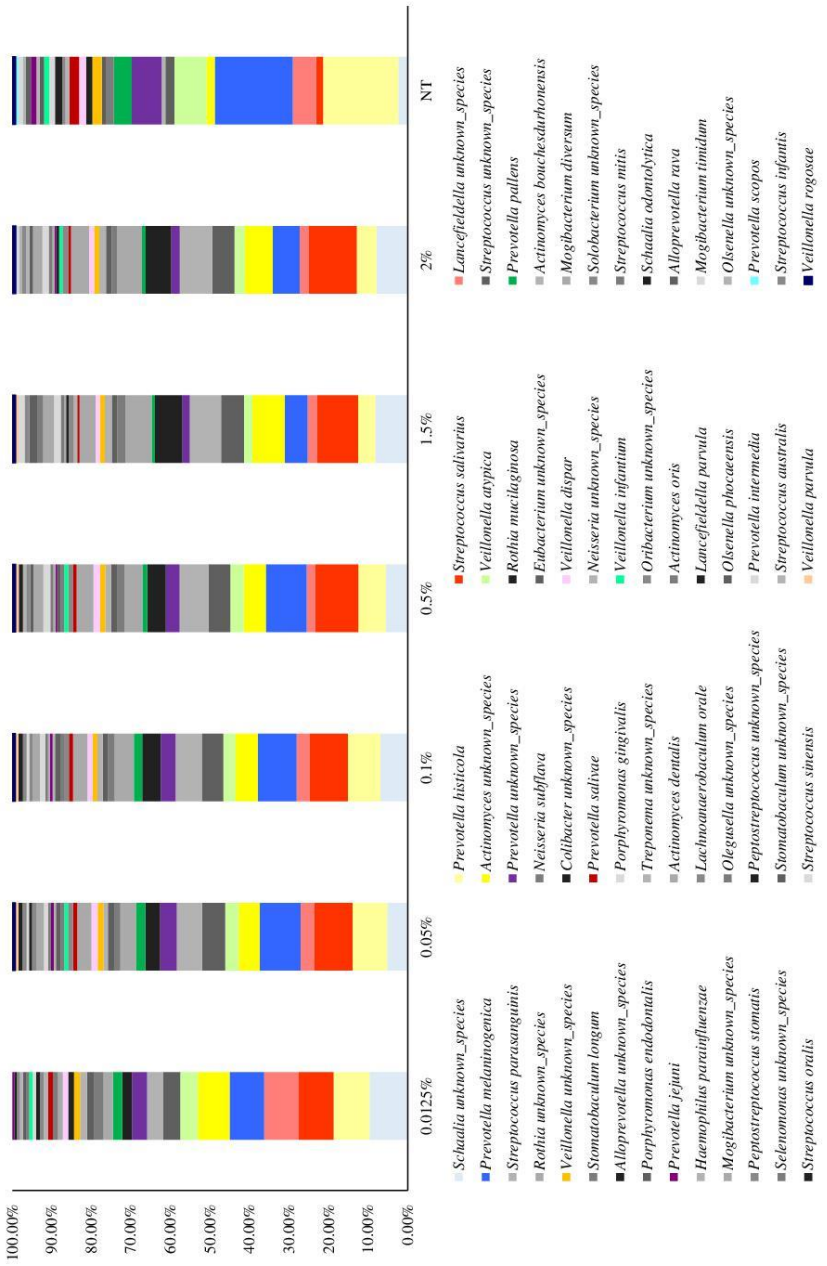


Figure S3

Figure S3. Taxonomic microbial profiling of the same sputum sample treated with different saponin amounts. From left to right, the color-coding order of the legend reflects average abundances from largest to smallest. Most representative bacterial species are highlighted with different colors. The untreated sample is reported as NT.

References

1. Procop GW. Molecular diagnostics for the detection and characterization of microbial pathogens. Clin Infect Dis. 2007 Sep 1;45 Suppl 2:S99-S111. PubMed PMID: 17683022. Epub 2007/08/19.
2. Nikkari S, Lopez FA, Lepp PW, Cieslak PR, Ladd-Wilson S, Passaro D, et al. Broad-range bacterial detection and the analysis of unexplained death and critical illness. Emerg Infect Dis. 2002 Feb;8(2):188-94. PubMed PMID: 11897072. Pubmed Central PMCID: PMC2732447. Epub 2002/03/19.
3. Clarridge JE, 3rd. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clinical microbiology reviews. 2004 Oct;17(4):840-62, table of contents. PubMed PMID: 15489351. Pubmed Central PMCID: 523561.
4. Cummings LA, Kurosawa K, Hoogestraat DR, SenGupta DJ, Candra F, Doyle M, et al. Clinical Next Generation Sequencing Outperforms Standard Microbiological Culture for Characterizing Polymicrobial Samples. Clin Chem. 2016 Nov;62(11):1465-73. PubMed PMID: 27624135. Epub 2016/10/30.
5. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun. 2016 Jan 22;469(4):967-77. PubMed PMID: 26718401. Pubmed Central PMCID: PMC4830092. Epub 2016/01/01.
6. Wensel CR, Pluznick JL, Salzberg SL, Sears CL. Next-generation sequencing: insights to advance clinical investigations of the microbiome. J Clin Invest. 2022 Apr 1;132(7). PubMed PMID: 35362479. Pubmed Central PMCID: PMC8970668. Epub 2022/04/02.
7. Feigelman R, Kahlert CR, Baty F, Rassouli F, Kleiner RL, Kohler P, et al. Sputum DNA sequencing in cystic fibrosis: non-invasive access to the lung microbiome and to pathogen details. Microbiome. 2017 Feb 10;5(1):20. PubMed PMID: 28187782. Pubmed Central PMCID: PMC5303297. Epub 2017/02/12.
8. Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome. 2018 Feb 27;6(1):42. PubMed PMID: 29482639. Pubmed Central PMCID: PMC5827986. Epub 2018/02/28.
9. Horz HP, Scheer S, Vianna ME, Conrads G. New methods for selective isolation of bacterial DNA from human clinical specimens. Anaerobe. 2010 Feb;16(1):47-53. PubMed PMID: 19463963. Epub 2009/05/26.
10. Ahannach S, Delanghe L, Spacova I, Wittouck S, Van Beeck W, De Boeck I, et al. Microbial enrichment and storage for metagenomics of vaginal, skin, and saliva samples. iScience. 2021 Nov 19;24(11):103306. PubMed PMID: 34765924. Pubmed Central PMCID: PMC8571498. Epub 2021/11/13.
11. Bruggeling CE, Garza DR, Achouiti S, Mes W, Dutilh BE, Boleij A. Optimized bacterial DNA isolation method for microbiome analysis of human tissues. Microbiologyopen. 2021 Jun;10(3):e1191. PubMed PMID: 34180607. Pubmed Central PMCID: PMC8208965. Epub 2021/06/29.
12. Goltsman DSA, Sun CL, Proctor DM, DiGiulio DB, Robaczewska A, Thomas BC, et al. Metagenomic analysis with strain-level resolution reveals fine-scale variation in the human pregnancy microbiome. Genome Res. 2018 Oct;28(10):1467-80. PubMed PMID: 30232199. Pubmed Central PMCID: PMC6169887. Epub 2018/09/21.
13. Byrd AL, Belkaid Y, Segre JA. The human skin microbiome. Nature reviews Microbiology. 2018 Mar;16(3):143-55. PubMed PMID: 29332945.
14. Malone M, Johani K, Jensen SO, Gosbell IB, Dickson HG, Hu H, et al. Next Generation DNA Sequencing of Tissues from Infected Diabetic Foot Ulcers. EBioMedicine. 2017 Jul;21:142-9. PubMed PMID: 28669650. Pubmed Central PMCID: PMC5514496. Epub 2017/07/04.
15. Thoendel M, Jeraldo PR, Greenwood-Quaintance KE, Yao JZ, Chia N, Hanssen AD, et al. Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. J Microbiol Methods. 2016 Aug;127:141-5. PubMed PMID: 27237775. Pubmed Central PMCID: PMC5752108. Epub 2016/05/31.

16. Hasan MR, Rawat A, Tang P, Jithesh PV, Thomas E, Tan R, et al. Depletion of Human DNA in Spiked Clinical Specimens for Improvement of Sensitivity of Pathogen Detection by Next-Generation Sequencing. J Clin Microbiol. 2016 Apr;54(4):919-27. PubMed PMID: 26763966. Pubmed Central PMCID: PMC4809942. Epub 2016/01/15.
17. Castelino M, Eyre S, Moat J, Fox G, Martin P, Ho P, et al. Optimisation of methods for bacterial skin microbiome investigation: primer selection and comparison of the 454 versus MiSeq platform. BMC Microbiol. 2017 Jan 21;17(1):23. PubMed PMID: 28109256. Pubmed Central PMCID: PMC5251215. Epub 2017/01/23.
18. Heravi FS, Zakrzewski M, Vickery K, Hu H. Host DNA depletion efficiency of microbiome DNA enrichment methods in infected tissue samples. J Microbiol Methods. 2020 Mar;170:105856. PubMed PMID: 32007505. Epub 2020/02/03.
19. Amar Y, Lagkouvardos I, Silva RL, Ishola OA, Foesel BU, Kublik S, et al. Pre-digest of unprotected DNA by Benzonase improves the representation of living skin bacteria and efficiently depletes host DNA. Microbiome. 2021 May 26;9(1):123. PubMed PMID: 34039428. Pubmed Central PMCID: PMC8157445. Epub 2021/05/28.
20. Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat Biotechnol. 2019 Jul;37(7):783-92. PubMed PMID: 31235920. Epub 2019/06/27.
21. Masud T, Kemp E. Corticosteroids in treatment of disseminated tuberculosis in patient with HIV infection. Br Med J (Clin Res Ed). 1988 Feb 13;296(6620):464-5. PubMed PMID: 3126861. Pubmed Central PMCID: PMC2545044. Epub 1988/02/13.
22. He Y, Fang K, Shi X, Yang D, Zhao L, Yu W, et al. Enhanced DNA and RNA pathogen detection via metagenomic sequencing in patients with pneumonia. J Transl Med. 2022 May 4;20(1):195. PubMed PMID: 35509078. Pubmed Central PMCID: PMC9066823. Epub 2022/05/06.
23. Khan MI, Ahhmed A, Shin JH, Baek JS, Kim MY, Kim JD. Green Tea Seed Isolated Saponins Exerts Antibacterial Effects against Various Strains of Gram Positive and Gram Negative Bacteria, a Comprehensive Study In Vitro and In Vivo. Evidence-based complementary and alternative medicine : eCAM. 2018;2018:3486106. PubMed PMID: 30598684. Pubmed Central PMCID: 6287149.
24. Milani C, Hevia A, Foroni E, Duranti S, Turroni F, Lugli GA, et al. Assessing the fecal microbiota: an optimized ion torrent 16S rRNA gene-based analysis protocol. PloS one. 2013;8(7):e68739. PubMed PMID: 23869230. Pubmed Central PMCID: 3711900.
25. Handschur M, Karlic H, Hertel C, Pfeilstocker M, Haslberger AG. Preanalytic removal of human DNA eliminates false signals in general 16S rDNA PCR monitoring of bacterial pathogens in blood. Comp Immunol Microbiol Infect Dis. 2009 May;32(3):207-19. PubMed PMID: 18261798. Epub 2008/02/12.
26. Milani C, Lugli GA, Duranti S, Turroni F, Mancabelli L, Ferrario C, et al. Bifidobacteria exhibit social behavior through carbohydrate resource sharing in the gut. Scientific reports. 2015 Oct 28;5:15782. PubMed PMID: 26506949. Pubmed Central PMCID: 4623478.
27. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754-60. PubMed PMID: 19451168. Pubmed Central PMCID: 2705234.
28. Milani C, Casey E, Lugli GA, Moore R, Kaczorowska J, Feehily C, et al. Tracing mother-infant transmission of bacteriophages by means of a novel analytical tool for shotgun metagenomic datasets: METAnnotatorX. Microbiome. 2018 Aug 20;6(1):145. PubMed PMID: 30126456. Pubmed Central PMCID: 6102903.
29. 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. 2021 Jun 29:e0058321. PubMed PMID: 34184911. Pubmed Central PMCID: 8269244.

30. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Gohl DM, Beckman KB, et al. Evaluating the Information Content of Shallow Shotgun Metagenomics. mSystems. 2018 Nov-Dec;3(6). PubMed PMID: 30443602. Pubmed Central PMCID: 6234283.
31. Mai-Prochnow A, Clauson M, Hong J, Murphy AB. Gram positive and Gram negative bacteria differ in their sensitivity to cold plasma. Sci Rep. 2016 Dec 9;6:38610. PubMed PMID: 27934958. Pubmed Central PMCID: PMC5146927. Epub 2016/12/10.
32. Lee JE, Jo SJ, Park KG, Suk HS, Ha SI, Shin JS, et al. Evaluation of modified saponin preparation method for the direct identification and antimicrobial susceptibility testing from positive blood culture. J Microbiol Methods. 2018 Nov;154:118-23. PubMed PMID: 30321566. Epub 2018/10/16.
33. Lupetti A, Barnini S, Morici P, Ghelardi E, Nibbering PH, Campa M. Saponin promotes rapid identification and antimicrobial susceptibility profiling of Gram-positive and Gram-negative bacteria in blood cultures with the Vitek 2 system. Eur J Clin Microbiol Infect Dis. 2013 Apr;32(4):493-502. PubMed PMID: 23114724. Epub 2012/11/02.
34. Ganda E, Beck KL, Haiminen N, Silverman JD, Kawas B, Cronk BD, et al. DNA Extraction and Host Depletion Methods Significantly Impact and Potentially Bias Bacterial Detection in a Biological Fluid. mSystems. 2021 Jun 29;6(3):e0061921. PubMed PMID: 34128697. Pubmed Central PMCID: PMC8574158. Epub 2021/06/16.
35. Marquet M, Zollkau J, Pastuschek J, Viehweger A, Schleussner E, Makarewicz O, et al. Evaluation of microbiome enrichment and host DNA depletion in human vaginal samples using Oxford Nanopore's adaptive sequencing. Scientific reports. 2022 Mar 7;12(1):4000. PubMed PMID: 35256725. Pubmed Central PMCID: 8901746.

Chapter 4

The Probiotic Identity Card: a novel 'probiogenomics' approach to investigate probiotic supplements

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 results of this chapter were published in Frontiers in Microbiology, 2022 Jan 21; 12:790881. doi: 10.3389/fmicb.2021.790881.

*These authors contributed equally.

Abstract

Probiotic bacteria are widely administered as dietary supplements and incorporated as active ingredients in a variety of functional foods due to their purported health-promoting features. Currently available probiotic products may have issues with regards to their formulation, such as insufficient levels of viable probiotic bacteria, complete lack of probiotic strains that are stated to be present in the product, and the presence of microbial contaminants. To avoid the distribution of such unsuitable or misleading products, we propose here a novel approach named Probiotic Identity Card (PIC), involving a combination of shotgun metagenomic sequencing and bacterial cell enumeration by flow cytometry. PIC was tested on 12 commercial probiotic supplements revealing several inconsistencies in the formulation of five such products based on their stated microbial composition and viability.

Keywords: Genomics, Metagenomics, Probiotics, Cell viability, Flow cytometry.

For Supplementary Materials see the article published in *Frontiers in Microbiology*.

Introduction

In 2009 a novel discipline called Probiogenomics was coined to provide insights into the diversity of probiotic bacteria aimed at revealing the molecular basis for their health-promoting activities (Ventura et al., 2009). In this context, the availability of probiotic genome sequences significantly expanded our understanding of the biology of these microorganisms (Ventura et al., 2012). Nonetheless, classical microbiological techniques are currently considered the gold standard for probiotic identification, classification, and enumeration (Chiron et al., 2018). However, these techniques are time-consuming and not always accurate when it comes to bacterial identification. In fact, most culture-based methods can only discriminate bacteria at the genus level and only detect microorganisms that can be cultivated (Chiron et al., 2018). In this context, several studies have described efforts to identify the microbial composition of commercial probiotic products that are sold to the US and European markets, encountering products lacking viable bacteria and/or with microbial compositions that deviate from the composition declared by the producers (Drago et al., 2010; Toscano et al., 2013). In recent years, next-generation sequencing technologies have enabled accurate evaluation of the relative abundance of (probiotic) microbes in a sample by targeting the 16S rRNA-encoding gene, thereby avoiding culture-dependent approaches (Morovic et al., 2016; Patro et al., 2016).

More recently, metagenomic sequencing has allowed compositional analysis of 10 probiotic supplements through 16S rRNA gene-associated sequencing and Whole Metagenome Shotgun (WMS) sequencing (Lugli et al., 2019). This analysis revealed inconsistencies of the bacterial presence in four out of 10 probiotic formulations assayed. Nonetheless, using the latter approach, enumeration of probiotic cells in each probiotic supplement was still missing, making it impossible to evaluate the viable count as previously identified by culture-based methods. Flow cytometry (FC) is used extensively in the field of microbiology to count bacteria and determine their viability and metabolic activities (Mudroňová, 2015; Chiron et al., 2018). Furthermore, it has recently been shown that FC is a valid analytical method to quantify lactic acid bacteria (Pane et al., 2018).

We here describe an novel analytic approach, named Probiotic Identity Card (PIC), which was initiated to improve the previously proposed Genetic Identity Card protocol (Lugli et al., 2019) through the use of FC assays to determine absolute abundance and viability of probiotic microorganisms in a given product/sample. In addition, gene-targeted metagenomic analyses involving the 16S rRNA gene and ITS profiling have been replaced by shotgun metagenomics, allowing probiotic classification at species level using a single sequencing methodology.

Results and Discussion

The PIC workflow

To perform a detailed microbial compositional assessment of probiotics, probiogenomics approaches were implemented involving next-generation sequencing and multiple FC assays. The workflow of this approach, schematically illustrated in Figure 1, consists of an initial step in which powder-based probiotic supplements are subjected to WMS sequencing. Sequenced DNA was then taxonomically classified at species level to reveal the relative abundance of each microorganism identified in the sample. At the same time, FC assays were performed using serial dilutions of the probiotic supplements, in order to enumerate bacterial cells and reveal their viability using dyes that distinguish live cells from dead cells based on cell membrane integrity. Following this, normalization of the WMS sequencing results was performed to estimate the absolute abundance of each viable probiotic strain within each sample assayed. These analyses were complemented by the enumeration of probiotic cells, and the evaluation of their viability. Another important step of the PIC protocol includes the probiotic genome sequence reconstruction based on WMS data obtained from the sequencing methodology, and the completeness of the assembled chromosomes was further validated using *in silico* programs aimed at identifying marker genes. Finally, where the probiotic formulation included microorganisms taxonomically classified as subspecies, the accuracy of the taxonomic identification was verified using a pangenome-based approach. In this context, a phylogenetic tree was built using multiple type strain sequences of the correlated subspecies.

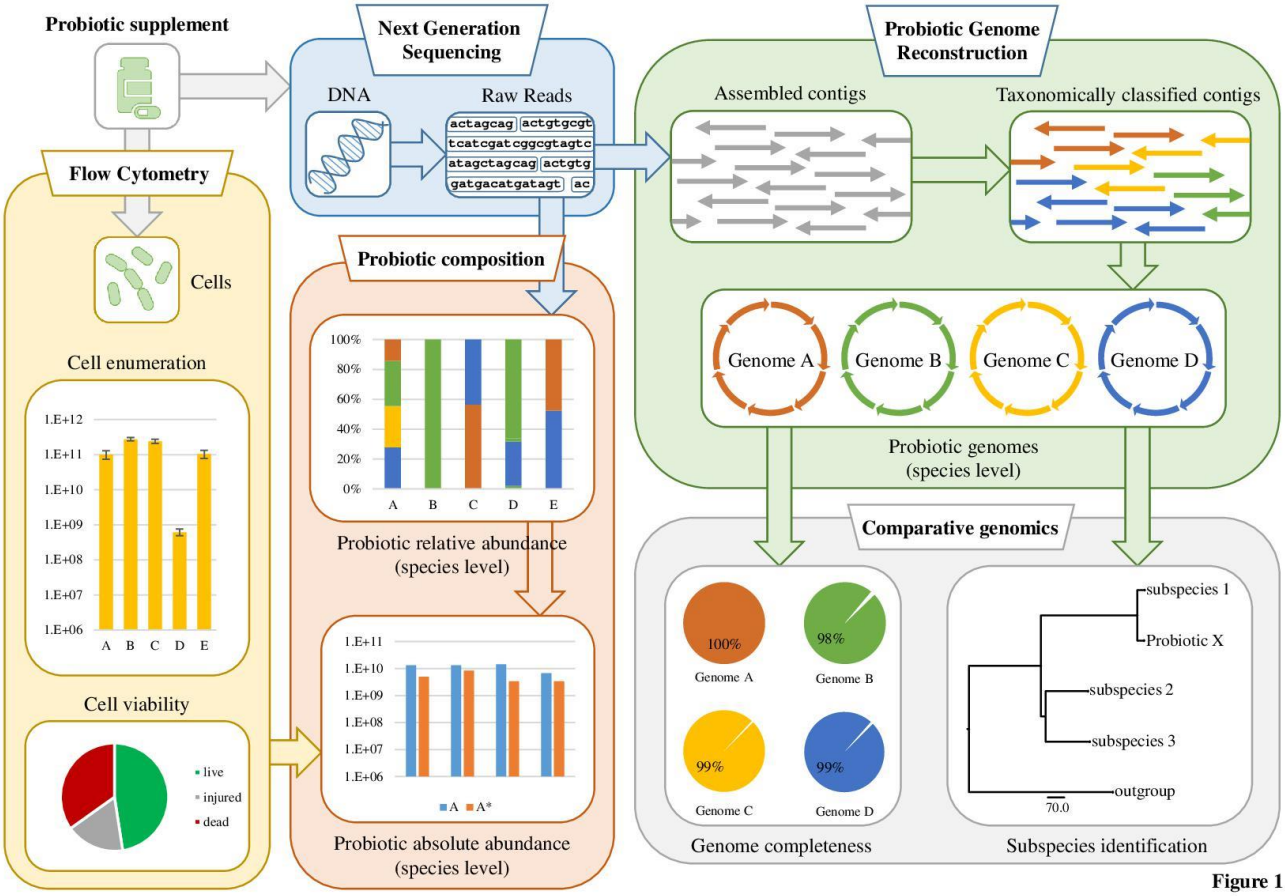


Figure 1

Figure 1. Schematic representation of the probiogenomics-based approach named Probiotic Identity Card (PIC). The methodology involves a whole metagenome shotgun (WMS) analysis followed by taxonomic classification of the reads and genomic reconstruction of the probiotic chromosomes. Then, two flow cytometry assays allow cell enumeration and viability, generating data which together with the WMS analysis are used to determine the integrity and quality of the probiotic supplement formulation.

Taxonomical dissection of probiotic supplements

Twelve powder-based probiotic supplements were selected and named A to L to retain anonymity of the products and their commercial origin (Table 1). WMS sequencing was performed to check the microbial composition stated on the packaging of the product. Sequencing outputs ranged from 0.5 to 5.7 million paired-end reads per sample (Table S1), allowing an accurate assessment of the probiotic species included in each supplement. In detail, the disparity of sequenced reads obtained from samples was directly proportional to the number of putative different probiotic strains harbored by each probiotic supplement (Table 1). Accordingly, the taxonomic classification of the total amount of 35 million reads allowed the identification of four species of *Bifidobacterium*, i.e., *Bifidobacterium animalis*, *Bifidobacterium bifidum*, *Bifidobacterium breve*, and *Bifidobacterium longum*, and eight species of *Lactobacillus*, i.e., *Lactobacillus acidophilus*, *Lactobacillus casei* (recently reclassified as *Lacticaseibacillus casei*), *Lactobacillus paracasei* (recently reclassified as *Lacticaseibacillus paracasei*), *Lactobacillus plantarum* (recently reclassified as *Lactiplantibacillus plantarum*), *Lactobacillus reuteri* (recently reclassified as *Limosilactobacillus reuteri*), *Lactobacillus rhamnosus* (recently reclassified as *Lactiplantibacillus rhamnosus*), *Lactobacillus salivarius* (recently reclassified as *Ligilactobacillus salivarius*), and *Lactobacillus zeae* (recently reclassified as *Lacticaseibacillus zeae*) (Fig. 2) (Zheng et al., 2020). Furthermore, *Bacillus coagulans*, *Enterococcus faecium*, *Saccharomyces cerevisiae*, and *Streptococcus thermophilus* species were also detected (Fig. 2).

Table 1. Probiotic data reported on the products

Code[#]	Probiotic species	N° species	CFU (~10^9)[#]
A	*Bifidobacterium animalis* subsp. *lactis* *Bifidobacterium breve* *Lactobacillus paracasei* *Lactobacillus plantarum*	4	20
B	*Lactobacillus casei*	1	24
C	*Lactobacillus reuteri* *Lactobacillus rhamnosus*	2	1.5
D	*Bacillus coagulans* *Bifidobacterium animalis* subsp. *lactis* *Lactobacillus acidophilus* *Lactobacillus casei*	4	2
E	*Bifidobacterium animalis* subsp. *lactis* *Bifidobacterium bifidum* *Bifidobacterium longum* subsp. *longum* *Lactobacillus acidophilus* *Lactobacillus casei* *Lactobacillus paracasei* *Lactobacillus plantarum* *Lactobacillus rhamnosus* *Lactobacillus salivarius*	9	1
F	*Bifidobacterium animalis* subsp. *lactis* *Lactobacillus rhamnosus*	2	4.5
G	*Enterococcus faecium* *Lactobacillus acidophilus* *Saccharomyces cerevisiae*	3	4
H	*Bifidobacterium animalis* subsp. *lactis* *Lactobacillus acidophilus* *Lactobacillus paracasei* *Lactobacillus plantarum*	4	70
I	*Bifidobacterium animalis* subsp. *lactis* *Bifidobacterium breve* *Lactobacillus acidophilus* *Lactobacillus paracasei* *Lactobacillus rhamnosus*	5	5.5
J	*Bifidobacterium breve* *Bifidobacterium longum* subsp. *infantis* *Lactobacillus acidophilus* *Lactobacillus casei* *Lactobacillus delbrueckii* subsp. *bulgaricus* *Lactobacillus rhamnosus* *Streptococcus thermophilus*	7	11
K	*Bifidobacterium animalis* subsp. *lactis* *Bifidobacterium breve* *Lactobacillus acidophilus* *Streptococcus thermophilus*	4	50
L	*Bifidobacterium longum* subsp. *infantis* *Lactobacillus acidophilus* *Lactobacillus reuteri* *Lactobacillus rhamnosus*	4	7

[#] Probiotic names and CFU of each strain are not reported to keep anonymity of probiotic supplements.

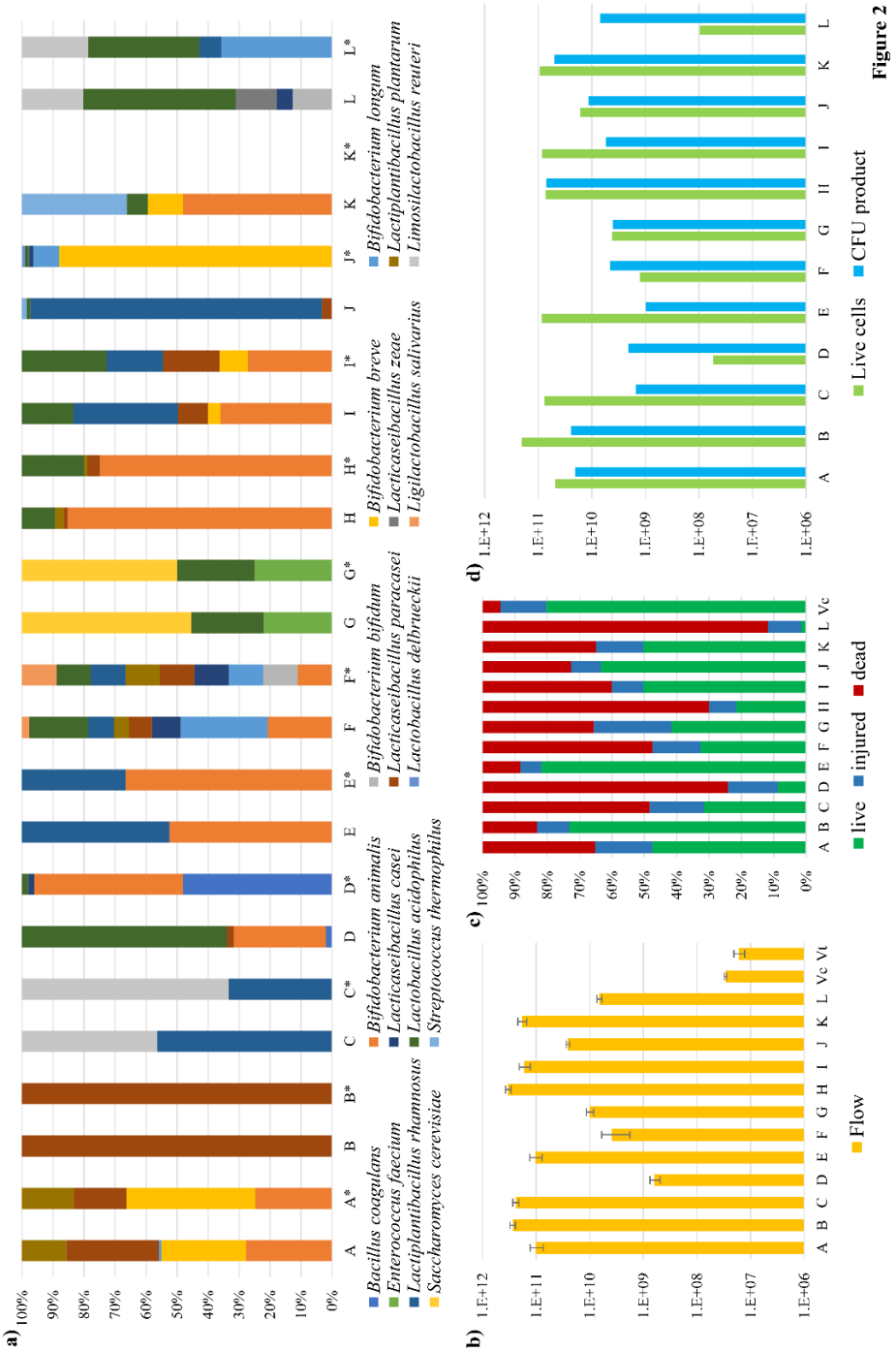


Figure 2. Microbial composition, quantification, and viability of probiotic supplements. Panel a displays the relative abundance of each microbial species identified in the analyzed probiotic samples. Each probiotic supplement is reported with an identification letter from A to L, which declared composition is listed in detail in Table 1 and reported in pillars marked with an asterisk. Panel b shows the quantification of the absolute number of microbial cells within each probiotic product. Pillars Vc and Vt report the cell enumeration data of a control sample by means of a flow cytometry assay and a counting chamber, respectively. Panel c depicts the percentage of viable, injured, and dead cells among individual samples. Finally, panel d exhibits the normalized number of viable cells (green) next to the value as stated by the producers (blue).

The analysis showed that probiotic products C, E, G, H, I, and K accurately reflect the bacterial composition as declared by the producer. In contrast, three probiotic supplements revealed that certain species were not present in the product, i.e., *B. bifidum* (product F), *B. breve*, *B. longum*, and *Lactobacillus delbrueckii* (product J), and *B. longum* and *L. rhamnosus* (product L) (Fig. 2). Furthermore, a more common inconsistency among the analyzed probiotic supplements was the presence of the *L. paracasei* taxon in samples B, D, and J, instead of the declared *L. casei* species (Table 1). This is a well-known issue since strains belonging to *L. casei* and *L. paracasei* are phenotypically and genotypically closely related (Huang et al., 2018). Furthermore, analysis of two probiotic supplements revealed the presence of additional bacteria not declared by the producers, i.e., bacteria belonging to *B. longum* (product A), and *B. bifidum*, *L. casei*, and *L. zeae* (product L) (Fig. 2).

Altogether, through WMS sequencing and subsequent taxonomic profiling of the sequences, we were able to precisely depict the microbial composition of the assessed samples (Fig. 2). If ignoring the *L. casei-paracasei* misclassification, probiotic supplements A and F revealed a minor contamination of *B. longum* (0.7 %) and the lack of a declared *B. bifidum* strain, respectively, while the observed composition of probiotic supplements J and L was shown to be inconsistent with their stated formulations. In particular, probiotic supplement L appeared to contain three unexpected microorganisms, of which *L. zeae* does not even belong to a generally accepted probiotic species.

Quantification and viability of probiotic strains

Even though WMS sequencing coupled with a bioinformatics approach does allow for a quick and accurate assessment of the microbial composition of probiotic supplements, the application of this approach can only unveil the relative abundance of microorganisms in each sample. Therefore, a flow cytometry (FC) assay was employed to enumerate the actual microbial cell number in each sample, thereby providing information on the absolute number of bacterial cells in the probiotic supplement. Based on the associated information leaflets of the analyzed probiotic supplements, the predicted

number of colony-forming units (CFU) ranged from one billion to 70 billion per capsule (Table 1). However, our FC analyses highlighted that bacterial cell numbers of these probiotic supplements ranged from 6.20 x 10^8 (stdev 1.38 x 10^8) in sample D to 3.34 x 10^{11} (stdev 3.79 x 10^{10}) in sample H (Table S2), highlighting that samples D and F contain cell numbers that are already lower than the viable CFU declared by the producer (Fig. 2). To validate the accuracy of this approach, an FC assay was further performed on a mock community encompassing two strains belonging to *B. bifidum* and *L. rhamnosus* species with a bacterial load of 1.67 x 10^7 CFU (stdev 3.82 x 10^6). Thus, cell enumeration by FC resulted in a number equaling 2.93 x 10^7 CFU (stdev 1.55 x 10^6), confirming the reliability of the FC assay established in this study (>0.05 Mann-Whitney U test) (Fig. 2).

A crucial feature of probiotic microorganisms is represented by their ability to interact with the human gut through the production of different metabolites (Lahtinen, 2012; Sánchez et al., 2017). Based on scientific literature published on this topic, health benefits conferred by viable probiotics are considered to be more prominent than those achieved by non-viable probiotics, also known as postbiotics (Żółkiewicz et al., 2020). Thus, the viability of microorganisms from each probiotic was also investigated by means of FC using dyes that distinguish live cells from dead cells based on cell membrane integrity. The percentage of viable cells with respect to the total load of bacterial cells as determined by FC ranged from 1.5 % in probiotic L to 81.9 % in probiotic E, with an average of 42 % of viable cells among probiotic supplements (Fig. 2). As a result of the analysis, products C, D, F, H, and L, revealed that the proportion of viable cells was less than 40 %, indicative of serious issues concerning the efficacy of these probiotic products. In addition, similar as described above for cell enumeration, an FC assay was performed on a mock community encompassing viable *B. bifidum* and *L. rhamnosus* taxa collected at the end of their exponential growth phase demonstrating the estimated presence of 5.5 % non-viable cells, thus validating the illustrated approach (Fig. 2).

Subsequently, FC data obtained for each probiotic supplement was further employed to normalize the total reads determined by WMS experiments according to a previously described method (Lugli et al., 2020), thereby allowing us to estimate the absolute abundance of each probiotic strain present in

the probiotic supplements assayed (Fig. 3). In detail, the composition of probiotic supplements C, E, G, H, and I were shown to be in near perfect agreement with what was declared by the producers for both the presence and absolute load, except for *L. paracasei* in sample H that exhibited an estimated cell number of 8.27×10^8 instead of 2.8×10^9 (Fig. 3). Furthermore, the microbial composition of probiotic A was also in very good correspondence with that stated in the accompanying information, except for an apparent contamination of *B. longum* (present at 3.24×10^8), which represents a relatively minor fraction when compared to the total estimated number of viable cells of 4.83×10^{10}. Since the information accompanying supplement K does not report the CFU of each strain, we did not include its absolute composition in this discussion. In contrast, probiotic supplements B, D, F, J, and L revealed serious discrepancies with respect to the absolute microbial content as stated by the manufacturers of these products (Fig. 3). In this context, we observed probiotic supplementations that were shown to contain much lower viable cells when compared to the number stated by the producers, in particular samples D (5.44×10^7 vs. 2.08×10^9) and L (9.62×10^7 vs. 7×10^9). Furthermore, we also noted formulations in which a single strain is numerically far more dominant compared to other strains in that same supplement as observed in sample J by *L. rhamnosus* (1.52×10^{10} vs. 2.2×10^7) (Fig. 3).

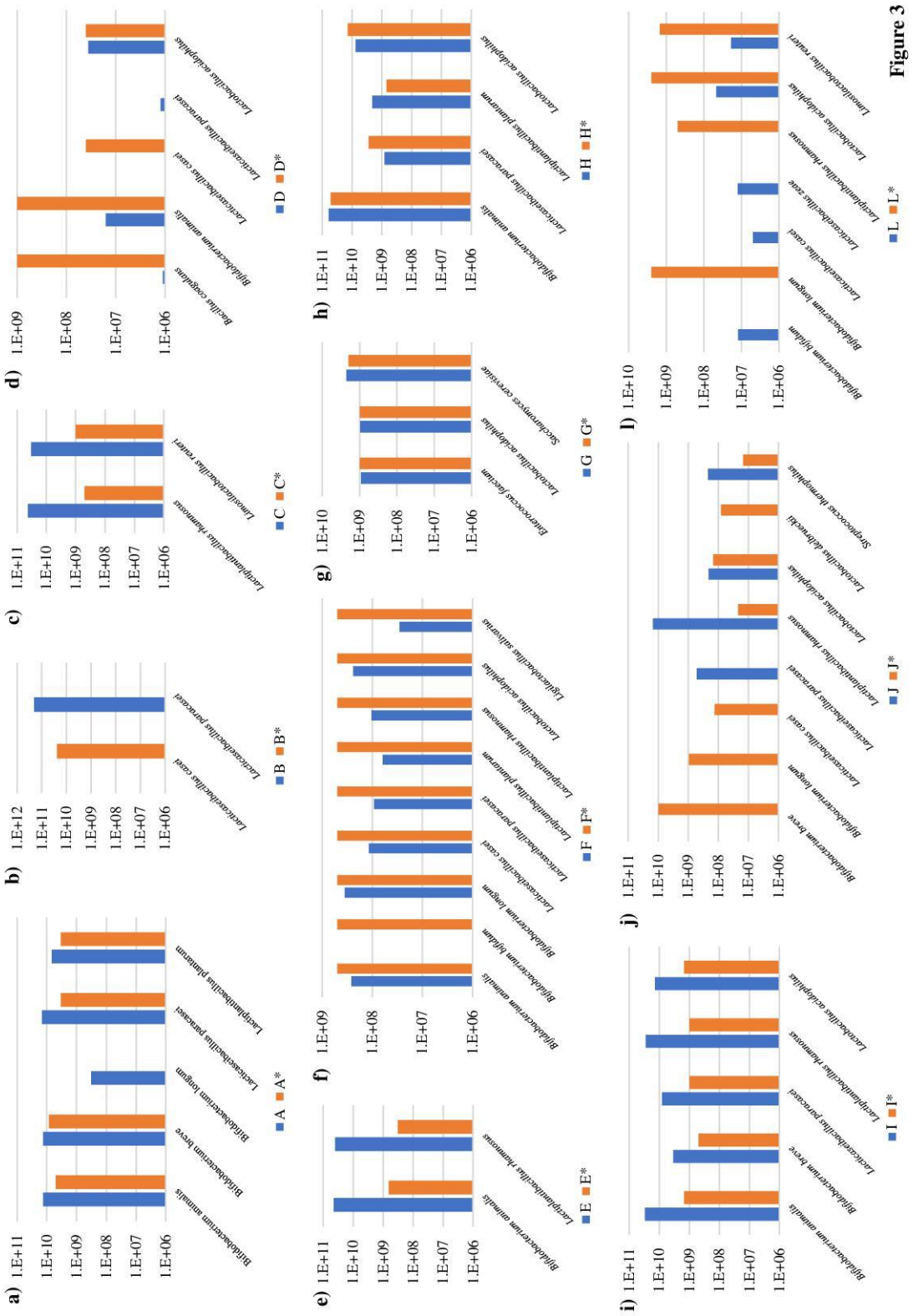


Figure 3. Absolute abundance values for microbes in a variety of probiotic supplements. Panel a to l exhibit the absolute abundance of each probiotic species among 11 supplements reported with an identification letter from A to L. Declared viability values of each microbial species is reported in pillars marked with an asterisk.

Genome reconstruction of probiotic strains

The high number of sequenced paired-end reads obtained for each probiotic product allowed us to perform metagenomic assembly to reconstruct the genome sequence of each probiotic microorganism. Assembled chromosomal sequences were taxonomically classified at the species level using a set of databases of validated reference genomes (Milani et al., 2021). Accordingly, shotgun sequencing data allowed the genome reconstruction of 46 chromosomal sequences encompassing strains belonging to *Bacillus coagulans*, *B. animalis*, *B. bifidum*, *B. breve*, *B. longum*, *E. faecium*, *L. acidophilus*, *L. casei*, *L. paracasei*, *L. plantarum*, *L. reuteri*, *L. rhamnosus*, *L. salivarius*, *L. zeae*, *Saccharomyces cerevisiae*, and *Streptococcus thermophilus* (Fig. 4). Likewise, reconstruction of such genome sequences revealed a microbial strain distribution across the 12 probiotic supplements identical to that predicted in the taxonomic classification of the short-read sequences (Fig. 2). In a similar fashion, assembled reads of product A unveiled a contig of 42Kb classified as *B. longum*, confirming the presence of putative contamination of this species in the formulation. Collected data further validated that products B, D, and F displayed some minor issues in the formulation represented by the misclassification of *L. paracasei* in *L. casei* and the absence of *B. bifidum* in sample F (Fig. 4). In contrast, products J and L lacked multiple strains and showed high contamination of other bacterial cells verifying the presence of *L. zeae* in sample L (Fig. 4). The (lack of) completeness of each reconstructed bacterial chromosome highlighted those strains previously estimated in low abundance within samples (Fig. 2), resulting in the partial genomic reconstruction of *L. acidophilus* (32 % in sample J), *L. paracasei* (45 % in sample H), *L. salivarius* (27 % in sample F), and *Streptococcus thermophilus* (36 % in sample J) (Fig. 4) (Table S3).

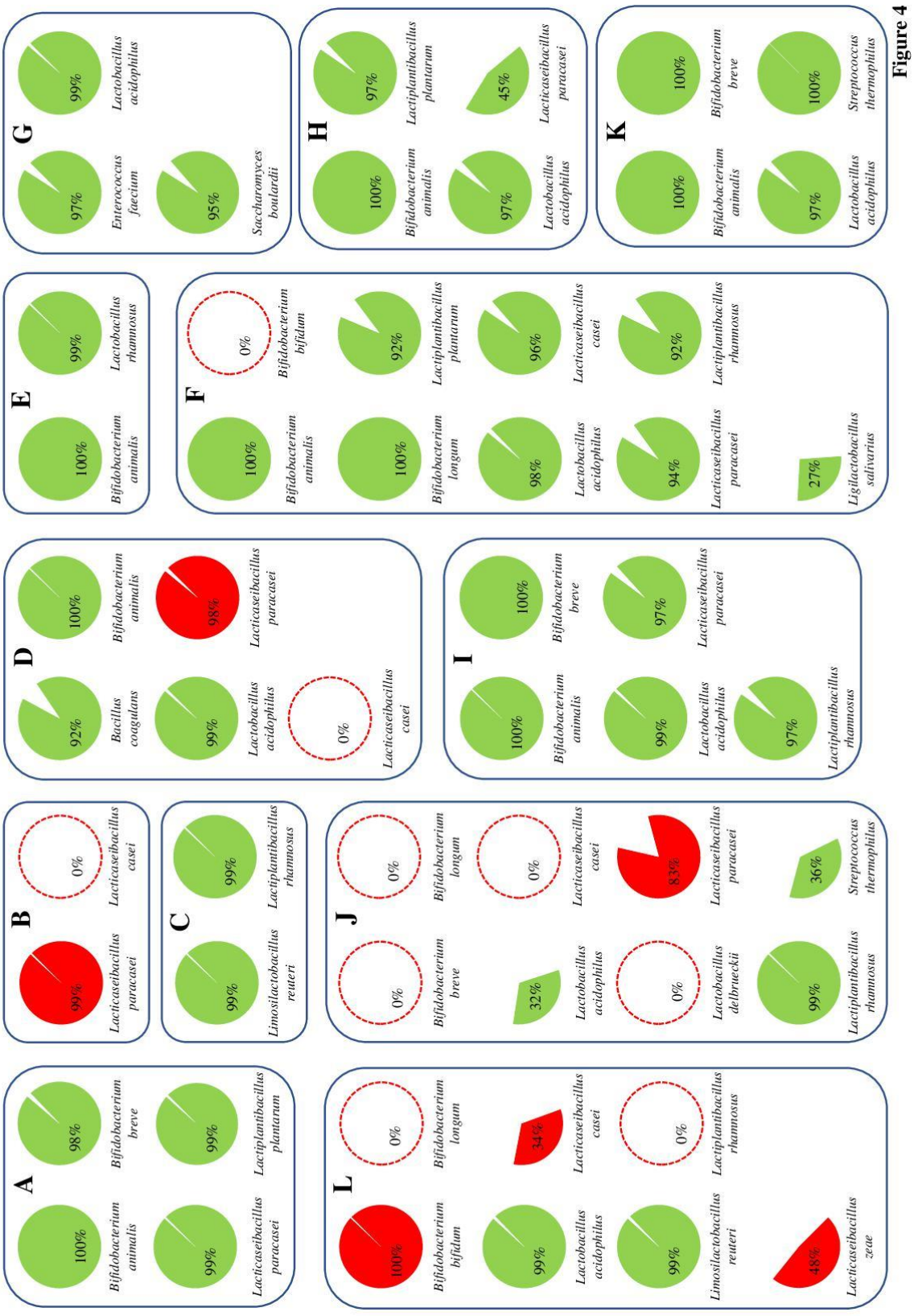


Figure 4

Figure 4. Assembled genomes from probiotic supplements. Reconstructed microbial genomes are represented by cake diagrams arranged in boxes for each of the 12 probiotic products. Missing microbes are reported as red dotted circles, while bacterial genomes not declared by the producers are highlighted in red. The percentage of each cake diagram corresponds to the completeness of the reconstructed chromosomes.

Since profiling of shotgun metagenomics data at the subspecies level is still very challenging, pangenome-based classification using genome sequences of related type strains have been used. In this context, nine out of the 12 assessed probiotic supplements were shown to contain bacteria that belong to *B. animalis* subsp. *lactis*, *B. longum* subsp. *infantis*, and *B. longum* subsp. *longum*. This analysis in the PIC workflow allowed us to build a pangenome-based phylogenetic tree to classify each strain at subspecies level (Fig. S1). Results highlighted that reconstructed *B. animalis* genomes in samples A, D, E, F, H, I, and K are highly related from a phylogenetic perspective, all belonging to the subspecies *B. animalis* subsp. *lactis*, while the reconstructed *B. longum* in sample F was shown to belong to the subspecies *B. longum* subsp. *longum*. Thus, using reference genomes, we were able to verify all probiotic strains classified as subspecies.

Conclusions

Using a combination of WMS sequencing and FC analyses, we characterized the microbial contents of commercial, powder-based, probiotic products, unveiling their presence as well as abundance and viability. The PIC pipeline described in this work improves the previously proposed Genetic Identity Card (Lugli et al., 2019), removing redundant sequencing experiments, such as 16S rRNA gene and ITS profiling, and allowing the precise enumeration of viable cells of each probiotic strain present in the probiotic supplement. Thus, the PIC approach can validate the probiotic formulation in terms of presence and absolute abundance of each probiotic microorganism, estimating the stated accuracy of the final product. Furthermore, no additional culturomic-based experiments are required, removing bias related to the differential grow capability of strains on different substrates. The PIC approach is also particularly suitable for dissecting the composition and verifying cell viability of probiotic supplements that encompass multiple probiotic strains (also known as mixes of probiotic bacteria) that are otherwise difficult to assess using standard culture-based approaches.

Using the PIC approach, five probiotic supplements out of 12, i.e., >40 %, reveal inconsistences in the formulations regarding what was declared, thus raising concerns about the current protocols applied by the internal quality checks of the probiotic supplement producer. Hopefully, molecular approaches such as the one described in this study will be established in the future by national agencies charged with quality control of probiotic products on the market.

Materials and Methods

Probiotic products selection

Twelve powder-based probiotic supplements were randomly selected from the Italian market to analyze probiotic products composed of single- and multi-strain microorganisms. Commercial names of collected products retrieved from pharmacies were renamed to observe anonymity of the producers. In detail, probiotic supplements were named with alphabetic letters, from A to L, and their microbial composition declared by the anonymous producers was listed in Table 1.

Microbial DNA extraction

Probiotic supplements were dissolved and homogenized thoroughly in Phosphate Buffer Solution (PBS; pH 6.5) to obtain the primary 1:10 dilution of each tested sample. Subsequently, one mL of each resuspended freeze-dried sample was subjected to chromosomal DNA extraction using the ZymoBIOMICS DNA Miniprep Kit (Zymo Research, D4300) following the manufacturer's instructions. Then, each probiotic supplement's DNA concentration and purity were investigated employing a Picodrop microtiter Spectrophotometer (Picodrop, Hinxton, UK).

Shotgun metagenomics sequencing

According to the manufacturer's instructions, DNA library preparation was performed using the Nextera XT DNA sample preparation kit (Illumina, San Diego, CA, USA). First, one ng input DNA from each probiotic supplement was used for the library preparation which underwent fragmentation,

adapter ligation, and amplification. Then, Illumina libraries were pooled equimolarly, denatured, and diluted to a concentration of 1.5 pM. Next, DNA sequencing was performed on a NextSeq 550 instrument (Illumina) using a 2X 150 bp Output sequencing Kit together with a deliberate spike-in of 1% PhiX control library.

Taxonomic classification of short sequenced reads

Sequenced paired-end reads of each probiotic supplement were subjected to a filtering step removing low-quality reads (minimum mean quality score 20, window size 5, quality threshold 25, and minimum length 100) using the fastq-mcf script (https://github.com/ExpressionAnalysis/ea-utils/blob/wiki/FastqMcf.md) to analyze high-quality sequenced data only. Then, an additional filtering step was performed to remove possible contaminating human DNA sequences from each sample through reads mapping employing the BWA aligner (Li and Durbin, 2009). Filtered reads were then collected and taxonomically classified through the METAnnotatorX2 pipeline (Milani et al., 2021), using a set of databases of reference genomes whose taxonomy was previously validated to maximize the accuracy of homology-based taxonomic classification of reads (Milani et al., 2021).

Genome reconstruction of probiotics through WMS sequencing

Filtered paired-end reads were subjected to whole metagenome assembly using Spades v3.15 (Prjibelski et al., 2020) with default parameters and the metagenomic flag option (--meta) together with k-mer sizes of 21, 33, 55, and 77. Reconstructed chromosomal contig sequences of probiotics were taxonomically classified against manually curated genome databases as reported above for the taxonomic classification of short sequenced reads (Milani et al., 2021). Overall, the METAnnotatorX2 pipeline was used to manage WMS data from read-filtering to taxonomic classification of the assembled contigs (Milani et al., 2018, 2021).

Flow cytometry analyses

From the initial 1:10 dilution of each probiotic supplement, five subsequent 10-fold serial dilutions were prepared in PBS. Then, one mL of bacterial cell dilution was stained with one µL mL^{-1} SYBR Green I (1:100 dilution in DMSO; Molecular Probes, Eugene, OR, USA) and incubated in the dark for 15 min before measurement. Count experiments were performed using an Attune NxT flow cytometer (ThermoFisher Scientific, Waltham, MA, USA) equipped with a blue laser set at 50mWand tuned to an excitation wavelength of 488 nm. Multiparametric analyses were performed on scattering signals, i.e., forward scatter (FSC) and side scatter (SSC), and SYBR Green I fluorescence was detected on the FL1 channel. Described analysis was performed in triplicate for each probiotic product. Cell debris was excluded from the acquisition analysis by a sample-specific FL1 threshold, and collected data were statistically analyzed with Attune NxT flow cytometer software. The precision of the enumeration method was determined on a mock community encompassing two strains belonging to *B. bifidum* and *L. rhamnosus* species, also coupled with a Thoma counting chamber calculation.

Microbial viability count

Two aliquots of one ml of bacterial cell dilution were harvested by centrifugation at 3,000 x *g* for 8 min. Cells were washed twice and resuspended in PBS. One of two aliquots of bacterial suspension was exposed to 70 % isopropyl alcohol for one hour to permeabilize cell membranes and cause cell death. Flow cytometry cell viability assay was carried out on both aliquots using the fluorescent stains SYTO9 (3.34 mM) and PI (20 mM) of LIVE/DEAD BacLight Bacterial Viability kit (ThermoFisher Scientific, Waltham, MA, USA), following the manufacturer's protocol. Specifically, 1.5 µl of a dye was added to the sample for the single staining assay, while for the double staining assay, 1.5 µl of both dyes was used. Immediately following staining, samples were incubated in the dark for 15 min at room temperature. For instrument parameter adjustment, single-colored controls were used, while non-stained cells were used as a background control. Cell viability assay was performed with an

Attune NxT flow cytometer (ThermoFisher Scientific, Waltham, MA, USA), and all data were analyzed with Attune NxT flow cytometer software. The precision of the viability count was determined on a mock community encompassing two strains belonging to *B. bifidum* and *L. rhamnosus* species.

Comparative genomics

The quality of reconstructed probiotic genomes was estimated for their completeness and contamination using CheckM v1.1.3 (Parks et al., 2015) and BUSCO v5 (Seppey et al., 2019). Then, pangenome calculations for subspecies identification of reconstructed probiotics were performed using the pangenome analysis pipeline PGAP (Zhao et al., 2012). Predicted proteomes were screened for orthologues between groups using BLAST analysis (cutoff E-value of $< 1 \times 10^{-5}$ and 50% identity across at least 80% of either protein sequence). The resulting output was clustered into protein families through MCL (graph theory-based Markov clustering algorithm) using the gene family method. Using this approach, phylogenetic trees were built, including genetic sequences of type strain retrieved from NCBI. The core genome trees were produced using the software FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Data availability

Shotgun metagenomics data are accessible through SRA study accession number PRJNA767508.

References

Chiron, C., Tompkins, T. A., and Burguière, P. (2018). Flow cytometry: a versatile technology for specific quantification and viability assessment of micro-organisms in multistrain probiotic products. *J. Appl. Microbiol.* 124, 572–584. doi:10.1111/jam.13666.

Drago, L., Rodighiero, V., Celeste, T., Rovetto, L., and de Vecchi, E. (2010). Microbiological evaluation of commercial probiotic products available in the USA in 2009. *J. Chemother.* 22, 373–377. doi:10.1179/joc.2010.22.6.373.

Huang, C. H., Li, S. W., Huang, L., and Watanabe, K. (2018). Identification and classification for the Lactobacillus casei group. *Front. Microbiol.* 9. doi:10.3389/fmicb.2018.01974.

Lahtinen, S. J. (2012). Probiotic viability – does it matter? *Microb. Ecol. Heal. Dis.* 23. doi:10.3402/mehd.v23i0.18567.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324.

Lugli, G. A., Duranti, S., Milani, C., Mancabelli, L., Turroni, F., Alessandri, G., et al. (2020). Investigating bifidobacteria and human milk oligosaccharide composition of lactating mothers. *FEMS Microbiol. Ecol.* 96. doi:10.1093/FEMSEC/FIAA049.

Lugli, G. A., Mangifesta, M., Mancabelli, L., Milani, C., Turroni, F., Viappiani, A., et al. (2019). Compositional assessment of bacterial communities in probiotic supplements by means of metagenomic techniques. *Int. J. Food Microbiol.* 294, 1–9. doi:10.1016/j.ijfoodmicro.2019.01.011.

Milani, C., Casey, E., Lugli, G. A., Moore, R., Kaczorowska, J., Feehily, C., et al. (2018). Tracing mother-infant transmission of bacteriophages by means of a novel analytical tool for shotgun metagenomic datasets: METAnnotatorX. *Microbiome* 6. doi:10.1186/s40168-018-0527-z.

Milani, C., Lugli, G. A., Fontana, F., Mancabelli, L., Alessandri, G., Longhi, G., et al. (2021). METAnnotatorX2: a Comprehensive Tool for Deep and Shallow Metagenomic Data Set Analyses. *mSystems* 6. doi:10.1128/msystems.00583-21.

Morovic, W., Hibberd, A. A., Zabel, B., Barrangou, R., and Stahl, B. (2016). Genotyping by PCR and high-throughput sequencing of commercial probiotic products reveals composition biases. *Front. Microbiol.* 7. doi:10.3389/fmicb.2016.01747.

Mudroňová, D. (2015). Flow cytometry as an auxiliary tool for the selection of probiotic bacteria. *Benef. Microbes* 6, 727–734. doi:10.3920/BM2014.0145.

Pane, M., Allesina, S., Amoruso, A., Nicola, S., Deidda, F., and Mogna, L. (2018). Flow Cytometry Evolution of Microbiological Methods for Probiotics Enumeration. *J. Clin. Gastroenterol.* 52, S41–S45. doi:10.1097/MCG.0000000000001057.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi:10.1101/gr.186072.114.

Patro, J. N., Ramachandran, P., Barnaba, T., Mammel, M. K., Lewis, J. L., and Elkins, C. A. (2016). Culture-Independent Metagenomic Surveillance of Commercially Available Probiotics with High-Throughput Next-Generation Sequencing. *mSphere* 1. doi:10.1128/msphere.00057-16.

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., and Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. *Curr. Protoc. Bioinforma.* 70. doi:10.1002/cpbi.102.

Sánchez, B., Delgado, S., Blanco-Míguez, A., Lourenço, A., Gueimonde, M., and Margolles, A. (2017). Probiotics, gut microbiota, and their influence on host health and disease. *Mol. Nutr. Food Res.* 61. doi:10.1002/mnfr.201600240.

Seppey, M., Manni, M., and Zdobnov, E. M. (2019). BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol. Biol.* 1962, 227–245. doi:10.1007/978-1-4939-9173-0_14.

Toscano, M., de Vecchi, E., Rodighiero, V., and Drago, L. (2013). Microbiological and genetic identification of some probiotics proposed for medical use in 2011. *J. Chemother.* 25, 156–161. doi:10.1179/1973947812Y.0000000068.

Ventura, M., O'Flaherty, S., Claesson, M. J., Turroni, F., Klaenhammer, T. R., van Sinderen, D., et al. (2009). Genome-scale analyses of health-promoting bacteria: Probiogenomics. *Nat. Rev. Microbiol.* 7, 61–71. doi:10.1038/nrmicro2047.

Ventura, M., Turroni, F., and van Sinderen, D. (2012). Probiogenomics as a tool to obtain genetic insights into adaptation of probiotic bacteria to the human gut. *Bioeng. Bugs* 3, 73–79. doi:10.4161/bbug.18540.

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

Zheng, J., Wittouck, S., Salvetti, E., Franz, C. M. A. P., Harris, H. M. B., Mattarelli, P., et al. (2020). A taxonomic note on the genus *Lactobacillus*: Description of 23 novel genera, emended description of the genus *Lactobacillus beijerinck* 1901, and union of *Lactobacillaceae* and *Leuconostocaceae*. *Int. J. Syst. Evol. Microbiol.* 70, 2782–2858. doi:10.1099/ijsem.0.004107.

Żółkiewicz, J., Marzec, A., Ruszczyński, M., and Feleszko, W. (2020). Postbiotics—a step beyond pre-and probiotics. *Nutrients* 12, 1–17. doi:10.3390/nu12082189.

Chapter 5

Tap water as a natural vehicle for microorganisms shaping the human gut microbiome

Longhi G*, Lugli G.A*, Mancabelli L, Alessandri G, Tarracchini C, Fontana F, Turroni F, Milani C, van Sinderen D, Ventura M.

The results of this chapter were published in Environmental Microbiology, 2022 Sep; 24(9):3912-3923. doi: 10.1111/1462-2920.15988.

*These authors contributed equally.

Summary

Fresh potable water is an indispensable drink which humans consume daily in substantial amounts. Nonetheless, very little is known about the composition of the microbial community inhabiting drinking water or its impact on our gut microbiota. In the current study, an exhaustive shotgun metagenomics analysis of the tap water microbiome highlighted the occurrence of a highly genetic biodiversity of the microbial communities residing in fresh water and the existence of a conserved core tap water microbiota largely represented by novel microbial species, representing microbial dark matter. Furthermore, genome reconstruction of this microbial dark matter from water samples unveiled homologous sequences present in the fecal microbiome of humans from various geographical locations. Accordingly, investigation of the fecal microbiota content of a subject that daily consumed tap water for three years provides proof for horizontal transmission and colonization of water bacteria in the human gut.

Originality-Significance Statement

Drinking water is a reservoir of microorganisms, the majority of which are not characterized due to the inability to cultivate them. Genome reconstruction of the water microbial dark matter allowed us to demonstrate that a large part of these microorganisms can be identified as a part of the human fecal microbiota. To our knowledge, this is the first work highlighting horizontal transmission and subsequent colonization of water microorganisms to the human gut, raising questions about the impact on human health of such microbial dark matter of water in modulating our gut microbiota.

Keywords: Metagenomics, Water, Microbiota, Microbial Dark Matter

For Supplementary Materials see the article published in *Environmental Microbiology*.

Introduction

Freshwater is estimated to represent about 2.5% of all water on Earth, while the remainder constitutes saltwater from seas and oceans. In developed countries, potable water is readily accessible as tap water and bottled natural mineral water, both of which are subject to strict safety regulations and very regular inspections (Eichler *et al.*, 2006). Conventional drinking water treatment plants perform filtration, sedimentation, disinfection, and flocculation, thereby allowing assessment of the microbial load (Dodd, 2012; Chao *et al.*, 2013; Loubet *et al.*, 2016). Nonetheless, some microorganisms may persist and proliferate in drinking water, with bacterial concentrations estimated to be around 10^6-10^8 cells per liter (Hammes *et al.*, 2008; Hong *et al.*, 2010). Thus, like other consumed items, water may represent a natural vehicle of microorganisms able to interact with the human gut and its microbiota (Dimidi *et al.*, 2019; Milani *et al.*, 2019).

Some studies have observed significant associations between tap water composition and human health (Bouchard *et al.*, 2011). In this context, differences were observed in the composition of the gut microbiota of mice that drank water from different sources, including tap water, highlighting an increase of clinically important taxa such as *Acinetobacter* and *Staphylococcus* in the feces and mucosa-adhered samples of animals (Dias *et al.*, 2018). Similar examples have also been reported in a human population context, given the recent evidence of bacterial spread from microbial biofilms present in drinking water distribution systems (Chan *et al.*, 2019) and investigations regarding the potential for tap water to influence human health mediated through the gut microbiota (Bowyer *et al.*, 2020).

Microbial populations in drinking water are challenging to assess because most of the bacteria present appear to be non-culturable (Loy *et al.*, 2005; França *et al.*, 2015) or are present in a viable-but-nonculturable state (Szewzyk *et al.*, 2000). For this reason, metagenomic sequencing approaches represent a well-established method to study the culturable and unculturable parts of the drinking water microbiota (Brumfield *et al.*, 2020; Sala-Comorera *et al.*, 2020). Furthermore, recent studies provide exciting insights into the composition of tap water microbiota by revealing that increased

urban development causes shifts in bacterial community composition of water (Simonin *et al.*, 2019). Thus, microbial community analyses can be helpful to diagnose environmental conditions as indicators of ecosystem health and water quality condition. In this context, the predominant bacteria that have been detected in drinking water are members of the phyla Actinobacteria and Proteobacteria, with the genera *Afipia*, *Bradyrhizobium*, and *Mycobacterium* being prevalent among tap water and drinking fountain samples (Brumfield *et al.*, 2020).

In the current study, we investigated the microbiota composition of tap water from the city pipeline of Parma, Italy, using a shallow metagenomics approach, which allows accurate taxonomic profiling of microbial communities from water samples down to species level. In addition, the metagenomics data was used in conjunction with data sets retrieved from the NCBI repository to describe the occurrence of a core tap water microbiota as well as the existence of specific microbial groups that are typical of the various geographical regions investigated. Within this context, we mapped the presence of a microbial taxon shared between the tap water microbiota and the gut microbiota of an individual daily drinking that water, suggesting horizontal transmission by these bacteria, which in turn appears to impact other resident members of the human gut microbiota.

Results and Discussion

Uncovering tap water microbial biodiversity. Sixteen water samples, named W001 to W016, were collected from public fountains (n = 12) and household taps (n = 4) throughout the Parma district in Italy to explore the microbial biodiversity of water samples across the city (Table 1). Shallow shotgun metagenomic sequencing was performed to identify the microorganisms that populate water samples at species level. Sequencing output constituted about one million paired-end reads, with an average of 64 thousand reads per sample (Table S1), thus allowing accurate assessment of the microbial species inhabiting each water sample (Hillmann *et al.*, 2018). This taxonomic survey revealed that just five species were present in particular samples at a relative abundance higher than 10 %, i.e., *Acidovorax delafieldii* (55 % in W012), *Aquabacterium commune* (26 % in W005), *Sphingomonas ursincola* (23 % in W011), *Sphingobium fluviale* (14 % in W015), and *Sphingomonas aerolata* (12 % in W009) (Fig. 1). As reported in literature, these species had previously been identified in biofilms of drinking water samples (Kalmbach *et al.*, 1999; Busse *et al.*, 2003; Morohoshi *et al.*, 2017). Moreover, unclassified members of 13 additional bacterial genera were identified with relative abundances ranging from 10 % to 39 %, highlighting the marked presence of as yet to be isolated and characterized bacterial species (Fig. 1). Interestingly, the major difference identified between samples collected from public fountains and household taps was represented by the presence of *Sphingomonas*, found as the most abundant taxa in nine out of 12 public fountains water samples (Table 1).

Table 1. Water samples of the Parma district.

Sample Name	Sample type	Sampling day	Most abundant species	SRA	Bioproject
W001	Tap water	22/06/2021	*Bradyrhizobium* sp.	SRR18015330	PRJNA806724
W002	Public Drinking Fountain Water	05/05/2021	*Sphingomonas* sp.	SRR18015329	PRJNA806724
W003	Public Drinking Fountain Water	05/05/2021	*Rothia* sp.	SRR18015322	PRJNA806724
W004	Tap water	28/05/2021	*Novosphingobium* sp.	SRR18015321	PRJNA806724
W005	Tap water	06/05/2021	*Aquabacterium commune*	SRR18015320	PRJNA806724
W006	Tap water	23/04/2021	*Nitrospira* sp.	SRR18015319	PRJNA806724
W007	Public Drinking Fountain Water	10/05/2021	*Sphingomonas* sp.	SRR18015318	PRJNA806724
W008	Public Drinking Fountain Water	07/05/2021	*Sphingomonas* sp.	SRR18015317	PRJNA806724
W009	Public Drinking Fountain Water	17/05/2021	*Sphingomonas aerolata*	SRR18015316	PRJNA806724
W010	Public Drinking Fountain Water	23/04/2021	*Sphingomonas* sp.	SRR18015315	PRJNA806724
W011	Public Drinking Fountain Water	17/05/2021	*Erythrobacter* sp.	SRR18015328	PRJNA806724
W012	Public Drinking Fountain Water	06/05/2021	*Acidovorax delafieldii*	SRR18015327	PRJNA806724
W013	Public Drinking Fountain Water	04/05/2021	*Sphingomonas* sp.	SRR18015326	PRJNA806724
W014	Public Drinking Fountain Water	04/05/2021	*Sphingomonas* sp.	SRR18015325	PRJNA806724
W015	Public Drinking Fountain Water	23/04/2021	*Sphingomonas* sp.	SRR18015324	PRJNA806724
W016	Public Drinking Fountain Water	07/05/2021	*Sphingomonas ursincola*	SRR18015323	PRJNA806724

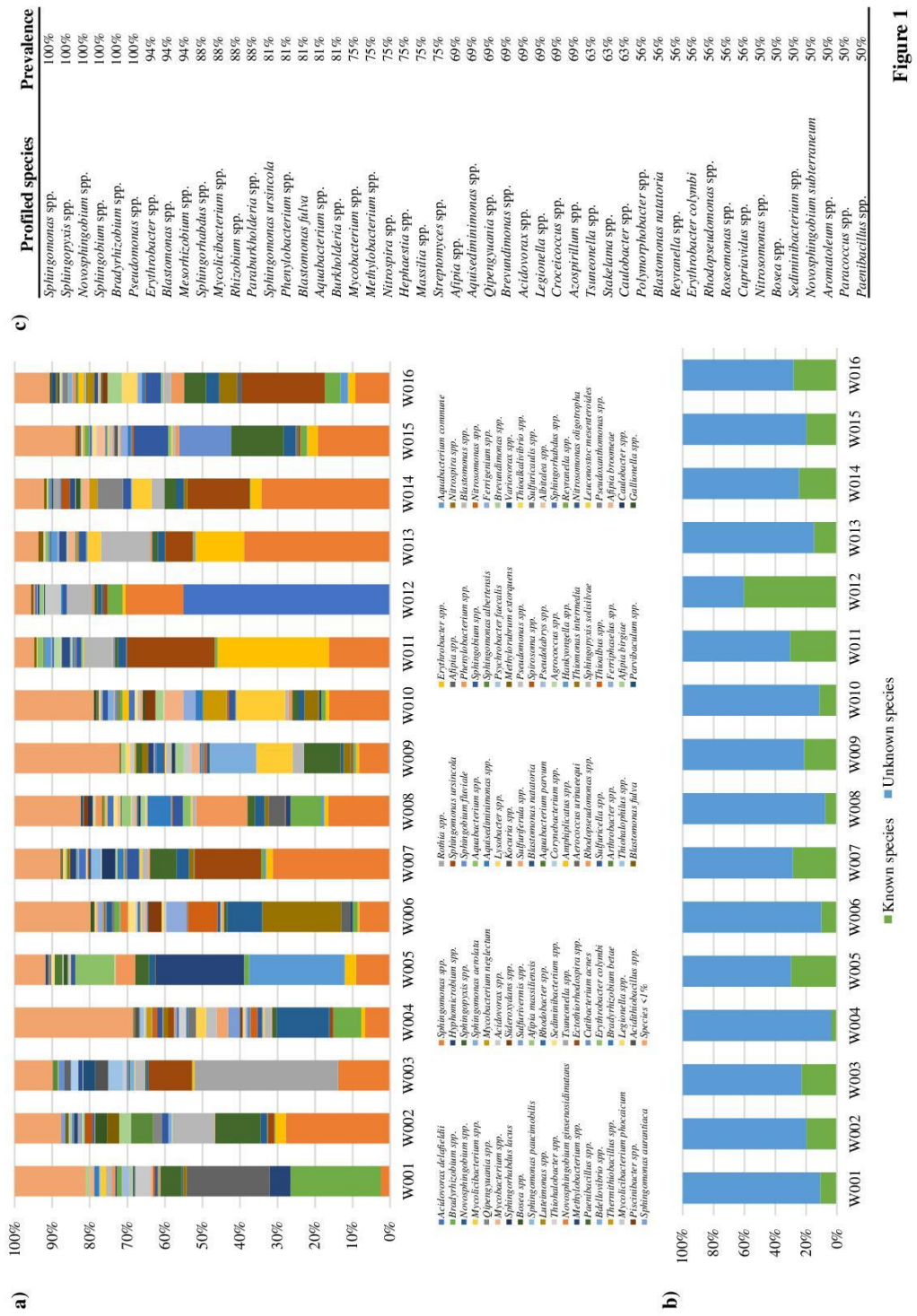


Figure 1

Figure 1. Microbial composition of 16 drinking water samples collected from different locations of the Parma district and delivered by the city water supply system. Panel a displays a histogram with the relative abundance of each microbial species identified in the analyzed samples. From left to right, the color-coding order of the legend reflects average abundances from largest to smallest, i.e., 55% *Acidovorax delafieldii* in sample W012, 38.9% *Sphingomonas* spp. in sample W013, and so on. Panel b shows the percentage of unknown and known bacterial species profiled, while Panel c depicts the prevalent bacterial taxa identified among all processed samples.

A flow cytometry (FC) assay was employed to enumerate microbial cells present in each water sample, thereby providing information on the absolute number of microbes. Based on the predicted number of colony-forming units (CFU), analyzed samples ranged from $2x10^3$ CFU/ml in W009 to $5.3x10^6$ CFU/ml in W001 (Fig. S1). Interestingly, when excluding sample W001, the average CFU/ml between water samples was $2.8\ x10^4$, highlighting W001 as an outlier of the analysis with a high level of bacterial cells probably due to downstream effects of the water distribution system (Fig. S1). Then, FC data obtained for each water sample was used to normalize the taxonomically classified reads obtained through shallow shotgun metagenomics according to a previously described method (Lugli *et al.*, 2020), thereby allowing an estimation of absolute abundance (Fig. S1). The findings showed a high CFU/ml number of *Bradyrhizobium* spp. and *Afipia* spp. present in sample W001, and *Rothia* spp. representing a microorganism present at high absolute abundance in sample W003, highlighting bacterial taxa which had also previously been identified in drinking water systems (Brumfield *et al.*, 2020).

Furthermore, taxonomic profiling at species level allowed us to identify the core taxonomic elements of the water microbiome, i.e., taxa that occur with the highest prevalence. Notably, in addition to *Sphingomonas ursincola* and *Blastomonas fulva*, which were determined to be the most prevalent species being detected in 13 out of 16 samples, taxonomic profiling also revealed DNA of unknown bacterial species attributable to 30 genera distributed between samples at a prevalence of >70 %. Remarkably, these later findings underscore that many members of the microbial drinking water population have as yet not been subject to any characterization (Fig. 1). Among the latter microbial dark matter, unknown members of six genera were identified in all 16 samples, i.e., *Bradyrhizobium*, *Novosphingobium*, *Pseudomonas*, *Sphingobium*, *Sphingomonas*, and *Sphingopyxis*. Intriguingly, while *Pseudomonas* is a ubiquitous environmental bacterial genus, members of the other five genera have previously been isolated from groundwater and drinking water (McAlister *et al.*, 2002; Yoon *et al.*, 2005; Sheu *et al.*, 2013; Singh *et al.*, 2015; Gulati and Ghosh, 2017). Altogether, these

metagenomic data sets highlight the substantial knowledge gap pertaining to drinking water-associated bacteria.

Meta-analysis of the water microbiome across the world. To validate the quality of our metagenomic analysis and compare the taxonomic profiles of water samples identified in the current study with those of other geographical locations, sequencing data of 119 drinking water samples were retrieved from nine metagenomic projects, named P02 to P10, which cover seven different countries, (Table S2). Taxonomic profiling at species level was performed applying the same pipeline and parameters used for samples collected as part of our own study, as reported above. While three samples were discarded due to low sequencing data (<20,000 DNA sequence reads) (Table S1), along with the P04 sample that was not suitable for statistical purposes, the remaining 115 water samples were analyzed together with the microbiome data of 16 samples retrieved in the Parma district.

Beta-diversity investigation represented through Principal Coordinate Analysis (PCoA) based on Bray-Curtis dissimilarity index allowed exploration of the water biome biodiversity as based on different studies (Fig. 2). Marked biodiversity was encountered between almost all projects (PERMANOVA p-value of < 0.05) except for P6 (Table S3). These data highlighted unique signatures among water sample microbiomes associated with the same country, possibly due to geographic and environmental factors such as temperature and pH. As similarly reported in P1 (Parma district project), an average of 16 % of the microbial DNA of water was classified at species level, while the remaining 84 % was attributable to unknown microbial taxa, thus representing a sizable part of the water microbial dark matter. Investigating the prevalence of bacterial taxa between samples of the same project, a limited number of taxa (between 13 and zero) were observed with a prevalence >80 %, indicating a high degree of microbial biodiversity and inter-site variability (Fig. 2). Notably, we included two longitudinal studies (P3 and P8) and observed a higher number of high prevalence taxa (13 and 11, respectively), indicating lower inter-sample biodiversity, when considering samples collected longitudinally from the same site (Dai *et al.*, 2018; Vosloo *et al.*, 2021). For example, only

Reyranella soli was identified with a high prevalence (83 %) in P5 samples, while a similar observation was made for *Blastomonas fulva* and *Sphingomonas ursincola* in the case of P1 samples, which showed a prevalence of 81 % for these species (Fig. 2).

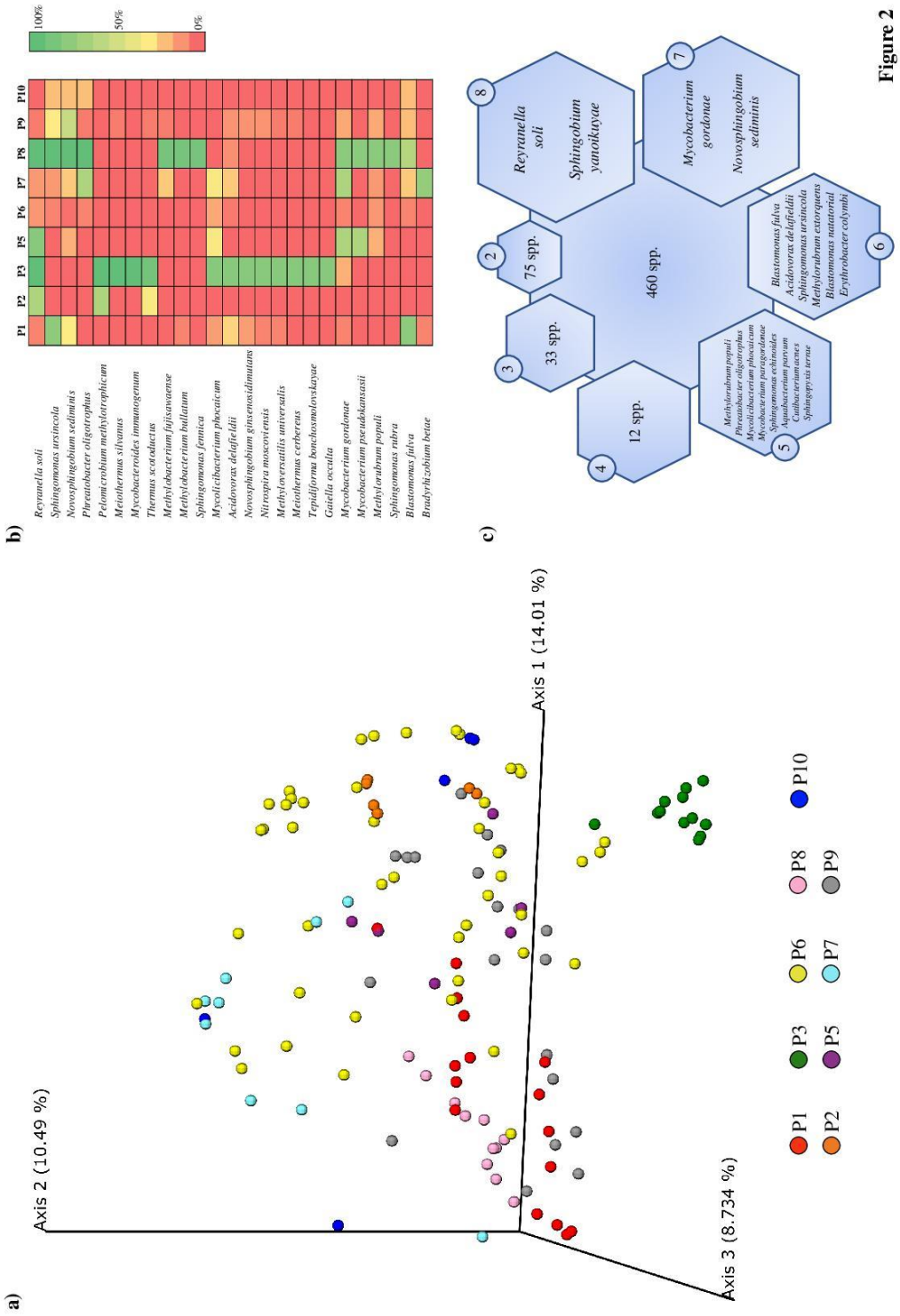


Figure 2. Microbial composition of 132 drinking water samples as obtained from NCBI. Panel a exhibits the principal coordinate analysis (PCoA), based on microbial distribution, represented in different colors by means of ten projects (P1 to P10) using the Bray–Curtis index. Panel b displays a heat map with the prevalence of major microbial players between projects. Panel c reports a schematic overview of the most prevalent bacterial species between projects. Species names and numbers in hexagons represent those taxa prevalent in samples from a number of projects written in circles. P4 data were excluded from the comparisons to avoid bias in the analyses' outcome.

Interestingly, 84 % of unclassified microbial DNA identified among water samples, previously referred as microbial dark matter (Rinke *et al.*, 2013), represented the actual core water microbiota (Table 2). In this context, following DNA filtering steps (see Materials and Methods), the microbial portion of a sample that was not classified at species level revealed unknown bacterial species highlighting the absence of a reference genome deposited in the NCBI repository. The five most prevalent microorganisms identified in the 115 drinking water samples have already been identified as major players in the 16 samples collected in the tap water of Parma district, i.e., members of the genera *Bradyrhizobium*, *Sphingomonas*, *Pseudomonas*, *Novosphingobium,* and *Sphingobium* (Table 2). Additionally, the DNA of species belonging to *Paraburkholderia*, *Burkholderia*, and *Mesorhizobium* was also identified in more than 90 % of the profiled waters (Table 2).

Table 2. Prevalence of profiled microorganisms between projects.

	P1	P2	P3	P5	P6	P7	P8	P9	P10	ALL	CORE
Bradyrhizobium spp.	100%	100%	100%	83%	100%	100%	100%	100%	100%	99%	8
Sphingomonas spp.	100%	100%	83%	83%	100%	100%	100%	100%	100%	98%	7
Pseudomonas spp.	100%	100%	100%	83%	96%	100%	100%	100%	100%	98%	7
Novosphingobium spp.	100%	100%	83%	83%	96%	100%	100%	100%	100%	96%	6
Paraburkholderia spp.	88%	100%	92%	67%	91%	100%	100%	100%	100%	93%	5
Sphingobium spp.	100%	100%	33%	83%	98%	100%	100%	100%	100%	92%	6
Burkholderia spp.	81%	100%	100%	50%	89%	100%	100%	100%	100%	92%	6
Mesorhizobium spp.	94%	100%	83%	67%	85%	100%	100%	95%	100%	90%	4
Massilia spp.	75%	100%	83%	67%	85%	100%	100%	100%	80%	88%	4
Cupriavidus spp.	56%	100%	100%	67%	87%	78%	100%	100%	100%	87%	5
Mycolicibacterium spp.	88%	100%	83%	67%	87%	56%	82%	80%	100%	83%	2
Methylobacterium spp.	75%	100%	25%	67%	91%	100%	100%	85%	100%	83%	4
Reyranella spp.	56%	100%	100%	100%	87%	100%	100%	60%	100%	83%	6
Sphingopyxis spp.	100%	67%	58%	50%	76%	100%	100%	100%	60%	83%	4
Roseomonas spp.	56%	100%	75%	50%	83%	100%	100%	90%	100%	83%	4
Rhizobium spp.	88%	100%	33%	67%	78%	100%	100%	100%	100%	83%	5
Streptomyces spp.	75%	100%	100%	67%	76%	100%	82%	85%	80%	83%	3
Azospirillum spp.	69%	100%	58%	50%	83%	100%	100%	85%	100%	82%	4
Aquabacterium spp.	81%	100%	92%	50%	83%	44%	91%	85%	60%	80%	1
Acidovorax spp.	69%	100%	92%	33%	80%	67%	100%	90%	60%	80%	2
Mycobacterium spp.	75%	67%	75%	67%	83%	56%	100%	85%	80%	80%	1
Variovorax spp.	44%	100%	92%	50%	80%	56%	100%	95%	100%	80%	3
Brevundimonas spp.	69%	100%	25%	33%	76%	100%	100%	100%	40%	76%	4
Phenylobacterium spp.	81%	100%	83%	50%	59%	100%	91%	90%	40%	75%	2
Bosea spp.	50%	100%	50%	67%	78%	100%	100%	75%	40%	74%	3
Ramlibacter spp.	44%	100%	75%	50%	80%	44%	100%	85%	60%	73%	2
Hydrogenophaga spp.	44%	100%	92%	33%	76%	33%	100%	85%	60%	72%	2
Achromobacter spp.	25%	100%	67%	33%	80%	67%	73%	85%	100%	70%	2
Caulobacter spp.	63%	100%	75%	33%	54%	100%	91%	85%	60%	70%	2
Paracoccus spp.	50%	100%	0%	50%	67%	100%	100%	70%	100%	67%	4
Afipia spp.	69%	100%	58%	50%	70%	100%	91%	20%	40%	64%	2
Rhodoferax spp.	13%	100%	67%	17%	65%	22%	100%	90%	60%	61%	2
Rhodopseudomonas spp.	56%	100%	75%	50%	61%	100%	64%	35%	60%	61%	2
Microvirga spp.	31%	100%	0%	67%	74%	100%	100%	40%	80%	61%	3
Nitrospira spp.	75%	100%	100%	17%	52%	0%	100%	55%	60%	61%	3
Hyphomicrobium spp.	31%	100%	8%	17%	80%	100%	100%	40%	40%	61%	3
Rhodoplanes spp.	31%	100%	58%	67%	70%	100%	91%	25%	40%	61%	2
Comamonas spp.	31%	100%	83%	17%	57%	22%	82%	90%	60%	61%	1
Polaromonas spp.	19%	100%	58%	17%	70%	22%	82%	80%	40%	59%	1
Methylorubrum spp.	31%	83%	0%	50%	63%	100%	100%	55%	40%	58%	2
Flavobacterium spp.	38%	33%	67%	50%	50%	11%	91%	90%	60%	56%	0
Noviherbaspirillum spp.	19%	100%	67%	17%	65%	33%	82%	50%	60%	55%	1
Erythrobacter spp.	94%	0%	0%	17%	33%	100%	100%	90%	40%	55%	2
Aromatoleum spp.	50%	100%	92%	50%	48%	44%	73%	45%	20%	55%	1
Thauera spp.	31%	100%	75%	50%	57%	11%	91%	45%	40%	54%	1
Nitrosomonas spp.	50%	100%	25%	17%	52%	0%	100%	70%	40%	52%	2
Legionella spp.	69%	50%	8%	17%	65%	11%	82%	50%	60%	52%	0
Paenibacillus spp.	50%	100%	0%	50%	57%	44%	55%	65%	60%	52%	1
Curvibacter spp.	19%	67%	67%	17%	46%	44%	100%	70%	40%	52%	1
Sphingorhabdus spp.	88%	0%	58%	0%	52%	11%	36%	80%	20%	51%	0
Lysobacter spp.	31%	100%	0%	17%	70%	33%	18%	70%	60%	51%	1

Members of the *Sphingomonas* genera are Gram-negative bacteria isolated from many different land and water habitats thanks to their ability to survive at low nutrient concentrations. More recently, the genus has been subdivided into different genera, in which we can find two other prevalent microbial groups reported above, i.e., *Novosphingobium* and *Sphingobium* (Takeuchi *et al.*, 2001). Conversely, members of the *Bradyrhizobium* and *Mesorhizobium* genera are Gram-negative nitrogen-fixing bacteria that occur either as free-living soil bacteria or that are found in symbiotic interaction within the roots of leguminous plants (Lorite *et al.*, 2018). Similarly, members of the *Paraburkholderia* genus are also nitrogen-fixing bacteria correlated with plant growth promotion, while members of the related genus *Burkholderia* can also be pathogens for humans, being the case for several species of the *Pseudomonas* genus. However, since these latter microorganisms are ubiquitously distributed in drinking water, they are probably not harmful to humans. Nevertheless, the presence of opportunistic pathogens in rainwater and tap water storage systems has already been discussed, showing the natural occurrences of *Pseudomonas aeruginosa*, *Legionella* spp., and *Mycobacterium* spp. (Zhang *et al.*, 2021). Thus, it would be of great interest to unveil the identified taxa genomic capability so as to assess and understand possible interactions with humans upon ingesting water containing such microbes.

Recently, High-throughput molecular analyses of microbiomes have been used as a tool to monitor the wellbeing of aquatic environments, involving cultural-independent analyses such as metagenomics, metatranscriptomics, metaproteomics, and metabolomics (Michán *et al.*, 2021). In the frame of this work, water microbiome profiling revealed a conserved microbial core represented by mostly unclassified and uncharacterized bacteria, which highlights a dearth of knowledge from a genomic and functionality perspective. Thus, microbiomes of waters should be investigated through culturomics experiments to gain access to such novel microbial species. In addition, since many microorganisms cannot be cultivated using standard procedures, deep metagenome sequencing can be performed to reconstruct unknown bacterial genomes.

Investigating the impact of tap water on the human microbiome. Recent literature has supported the notion that ingestion of foods populated by a specific microbiota facilitates transmission and subsequent colonization of these microorganisms in the human gut (Hehemann *et al.*, 2010; Makki *et al.*, 2018; Milani *et al.*, 2019). Since every person consumes about two liters of water per day, we explored the influence of water consumption on shaping the composition of the human gut microbiota. Therefore, bacterial DNA sequences collected from the 115 drinking water samples belonging to P02-P10 were subjected to genome reconstruction to gather chromosomal fragments of unknown bacterial taxa. In total, 2.9 Gigabases of DNA sequences belonging to unknown bacteria were assembled in this manner. Nonetheless, we decided to investigate only those bacteria identified with the highest prevalence in water samples of P1 and the other projects (Table 2) represented by 622 Megabases of as yet unclassified members of the genera *Bradyrhizobium*, *Burkholderia*, *Mesorhizobium*, *Novosphingobium*, *Paraburkholderia*, *Pseudomonas*, *Reyranella*, *Sphingobium*, *Sphingomonas*, and *Sphingopyxis*.

To trace the presence of the assembled microbial genomes, we included in this study shotgun metagenomic data of 196 human fecal samples retrieved from the NCBI repository (Table S2). Selection of fecal samples was performed so as to equally cover countries from which the drinking water was collected (Table S2). Then, the total amount of microbial DNA retrieved from the latter samples was used to investigate the presence of bacterial dark matter reconstructed through shotgun metagenomic assemblies. DNA mapping was performed with high sensitivity and high specificity (see Materials and Methods), identifying DNA of the reconstructed bacteria from the water microbiome in 46 of the analyzed human fecal samples (Fig. 3). DNA sequences corresponding to unclassified bacteria were distributed across the three analyzed continents, with a higher predominance of sequences belonging to *Sphingomonas* and *Reyranella*, both identified in 43 fecal samples, followed by *Novosphingobium* in 37 fecal samples. Following the latter microorganisms, the DNA of *Sphingopyxis*, *Pseudomonas*, and *Bradyrhizobium* was identified in 27 to 25 samples, while the remaining four genera were identified in 16 to 14 fecal samples (Fig. 3). Altogether, these

data indicate that the DNA of microorganisms belonging to the core microbiota of drinking water samples can also be detected in human fecal samples. Overall, these findings suggest that such tap water-associated microorganisms contribute to the human gut microbiota composition. A further interesting aspect awaiting to be investigated will be how many of these water-associated microorganisms found in the gut of humans were transmitted through direct ingestion of water or to the extensive use of tap water in watering plants for consumption.

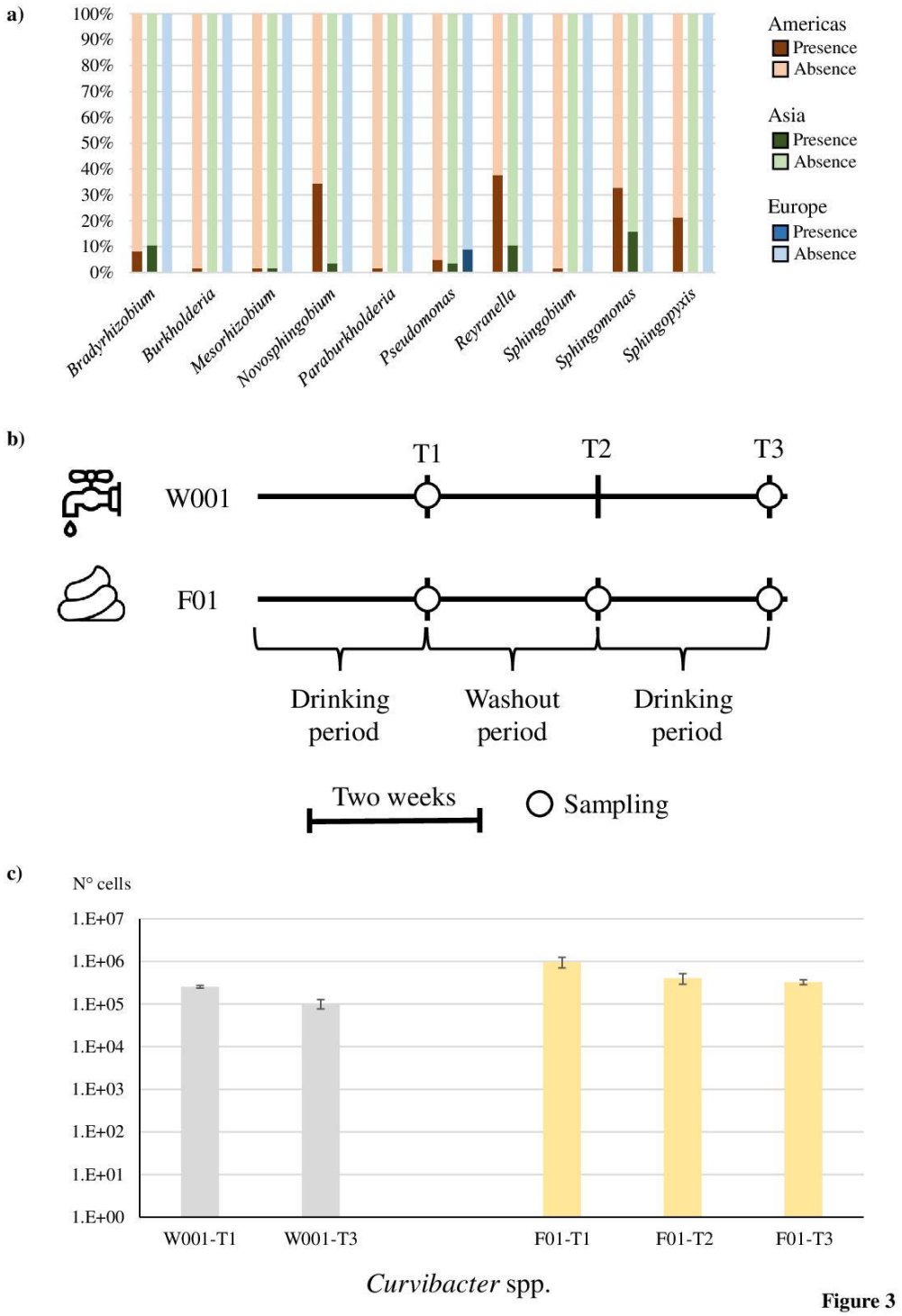


Figure 3. Investigation of the colonization of drinking water bacteria in humans. Panel a shows the percentage of human fecal samples in which the ten core microorganisms associated with drinking water were identified through read mapping. Panel b represents a schematic representation of the pilot experiment with drinking and sampling periods (T1, T2, and T3). Panel c illustrates the qPCR data of *Curvibacter* spp. in tap water (W001-T1 and W001-T3) and fecal samples of the subject (F01-T1, F01-T2, and F01-T3).

Transmission of microorganisms from tap water to the human gut. The *in silico*-based findings indicate the occurrence of microbial DNA belonging to the core tap water microbiota in various human fecal microbiomes. This prompted us to investigate this potential novel route for shaping the human gut microbiota. Thus, we examined the gut microbiota composition, together with the corresponding tap water microbiota, of a subject who daily consumed tap water for the past three years. Specifically, tap water consumed by the subject had been profiled in the context of this work (corresponding to W001) and subjected to metagenome assembly of its microbial DNA content. Assembled bacterial DNA resulted in 104 Kb distributed in 40 contigs, reflecting portions of chromosomes of those microorganisms present in relatively high abundance within the sample. Notably, assembled contigs, which taxonomically were predicted to belong to putative unknown species of the genus *Bradyrhizobium*, *Curvibacter*, and *Sphingobium*, were used to design primer pairs for quantitative real-time PCR (qPCR) investigation.

Fecal sample collection was accomplished through three time points aimed at investigating the microbiota of the subject during the consumption of W001 (T1 and T3) and after a period of washout of two weeks, during which the subject drank only bottled water (T2) (Fig. 3). Notably, bottled water microbiota was analyzed, revealing a negligible amount of microbial DNA, thus corroborating the microbiologically sterility of the administered water between T1 and T2. Due to the high CFU/ml identified in sample W001 at T1 ($5.3x10^6$), an additional (FC) assay was performed at T3 of the same tap water highlighting a consistent quantification of microbial cells ($4.8x10^6$). Remarkably, employing the *Bradyrhizobium* and *Sphingobium* strain-specific primers, qPCR assays on the fecal samples resulted in quantifying DNA below the detection limit. Instead, *Curvibacter* DNA was identified at each time point, indicating apparent colonization of this *Curvibacter* species in the subject's gut as it was detected even after two weeks following the start of the washout period (Fig. 3). Notably, even if these data have been obtained from a single individual, they suggest that some microbial species residing in tap water can survive to the human colonic tract and colonize and persist

in the gut of their human host. A clinical trial involving a substantial number of subjects will need to be executed to validate these findings.

Conclusions

Tap water is considered a food (Pilot, 2012), providing essential elements to our body, which are vital for our lives. However, tap water can also be a reservoir of microorganisms that, once ingested, may colonize our intestine, influence the gut microbiota and be responsible for different metabolic activities associated with human health. In the current study, we were interested in assessing the notion that water is not only crucial for our nutritional and physiological requirements but may also be important as a delivery vehicle of microorganisms to the gut. Notably, and in contrast to most consumed foods, the microbial community composition of tap water is not very well studied (Sala-Comorera *et al.*, 2020). Here, we clearly show that a large part of the microorganisms present in water is represented by as yet to be characterized bacteria, thus representing constituting a lot of microbial dark matter. These findings should prompt dedicated investigations on these bacteria, aimed at isolation, cultivation and subsequent dissection of their biological features. In fact, as outlined in our study, a large number of these putative novel bacterial taxa, which make up part of the core tap water microbiota, are also identified as part of the human fecal microbiota representing various different metagenomes and geographical regions. Thus, this co-sharing scenario of members of the core tap water microbiota and the human gut microbiota may impact on human health through modulation of the gut microbiota. This may therefore represent an intriguing and novel scenario that warrants further careful exploration. Here, our findings indicate that horizontal transmission and subsequent colonization of microorganisms from tap water to the human gut is possible. However, a clinical trial encompassing a more extensive set of individuals drinking tap water encompassing different microbiota and involving the isolation of the microorganisms using culturomics approaches needs to be performed in order to corroborate our data.

Experimental Procedures

Tap water samples and sampling conditions. Sixteen tap water samples, including public fountains and household taps, were randomly selected from different locations and distribution systems encompassing various parts of Parma town and its territory. To ensure that the collected samples are representative of consumed water, at least five liters of water was directly collected from the tap, keeping at a safe distance from the faucet and letting some flow down before directly flushing the water into sterile bottles to minimize any contaminations. Water samples were transported to the laboratory and kept at 4°C for further analysis.

Microbial DNA extraction. For bacterial DNA extraction, five liters of a given water sample was filtered through 0.45 μm pore size hydrophilic mixed cellulose esters (Pall Corporation, Port Washington, NY, USA). Filters were placed in standard Petri dishes and were cut into small pieces to ensure total sterility. DNA was extracted from the filters using the ZymoBIOMICS DNA Miniprep Kit (Zymo Research, D4300) following the manufacturer's instructions. Then, each tap water sample's DNA concentration and purity were investigated employing a Picodrop microtiter Spectrophotometer (Picodrop, Hinxton, UK).

Shallow shotgun sequencing. According to the manufacturer's instructions, DNA library preparation was performed using the Nextera XT DNA sample preparation kit (Illumina, San Diego, CA, USA). First, one ng input DNA from each sample was used for the library preparation, which underwent fragmentation, adapter ligation, and amplification. Then, Illumina libraries were pooled equimolarly, denatured, and diluted to a concentration of 1.5 pM. Next, DNA sequencing was performed on a MiSeq instrument (Illumina) using a 2X 250 bp Output sequencing Kit together with a deliberate spike-in of 1% PhiX control library.

Short read taxonomic classification. Sequenced paired-end reads of each water sample were subjected to a filtering step removing low-quality reads (minimum mean quality score 20, window size 5, quality threshold 25, and minimum length 100) using the fastq-mcf script (https://github.com/ExpressionAnalysis/ea-utils/blob/wiki/FastqMcf.md) to analyze high-quality sequenced data only. Then, an additional filtering step was performed to remove possible contaminating human DNA sequences from each sample through reads mapping employing the BWA aligner (Li and Durbin, 2009). Filtered reads were then collected and taxonomically classified through the METAnnotatorX2 pipeline (Milani *et al.*, 2021), using a set of databases of reference genomes whose taxonomy was previously validated to maximize the accuracy of homology-based taxonomic classification of reads (Milani *et al.*, 2021).

Metagenome assembly. Filtered reads were subjected to whole metagenome assembly using Spades v3.15 (Wedemeyer *et al.*, 2017) with default parameters and the metagenomic flag option (-meta) together with k-mer sizes of 21, 33, 55, and 77. As mentioned above, for the short reads, reconstructed contig sequences were taxonomically classified based on their sequence identity using megablast against the same RefSeq database (Chen *et al.*, 2015). ORFs of each assembled genome were predicted with Prodigal (Hyatt *et al.*, 2010) and annotated utilizing the MEGAnnotator pipeline (Lugli *et al.*, 2016). In all, the METAnnotatorX2 pipeline was employed for various purposes, from read filtering to taxonomic classification of the assembled contigs (Milani *et al.*, 2018, 2021).

Flow cytometry analysis. The samples for flow cytometry were collected in sterilized screw tap tubes (Sarsted) and were transported to the laboratory within one hour of collection, temporarily stored at 4°C, and measured within a few hours after collection. Then, one mL of water sample was stained with one µL mL^{-1} SYBR Green I (1:100 dilution in DMSO; Molecular Probes, Eugene, OR, USA) and incubated in the dark for 15 min before measurement. Count experiments were performed using an Attune NxT flow cytometer (ThermoFisher Scientific, Waltham, MA, USA) equipped with

a blue laser set at 50mW and tuned to an excitation wavelength of 488 nm. Multiparametric analyses were performed on scattering signals, i.e., forward scatter (FSC) and side scatter (SSC), and SYBR Green I fluorescence was detected on the FL1 channel. The detection limit was determined experimentally by filtering one aliquot of water sample and one of Attune Focusing Fluid 1X through 0.20 µm pore size hydrophilic mixed cellulose esters (Pall Corporation, Port Washington, NY, USA). Then one mL of each sample was stained with one µL mL^{-1} SYBR Green I as mentioned above. Cell debris was excluded from the acquisition analysis by a sample-specific FL1 threshold, and collected data were statistically analyzed with Attune NxT flow cytometer software.

DNA Mapping. Microbial DNA retrieved from 197 human fecal samples were aligned to the reconstructed chromosomal portions of unknown water bacteria to evaluate the presence of water microorganisms in the gut of humans. The Bowtie2 program was used to align the DNA sequences through multiple-hit mapping and a "very sensitive" policy (Langdon, 2015). The mapping was performed using a minimum score threshold function (–score-min C,-13,0) to limit reads of arbitrary length to one mismatch and retain those matches with at least 99% full-length identity. The SAMtools software package (Danecek *et al.*, 2021) was then used to count the mapped reads among each bacterial taxon, rejecting hits with less than ten reads to achieve a consistent output.

Experimental design. The experiment involves a healthy adult male who daily drank tap water corresponding to sample W001 for the last three years. The objective was to collect fecal samples before and after the two weeks of the washout period. The first fecal sample collection was performed before the washout (F01-T1) to identify bacteria introduced by the consumption of tap water. Then, we collected fecal samples at T2 and T3 to cover the end of the washout and the restoration of W001 administration. Fecal samples were stored at -80°C until use. Concomitantly, W001 was collected at T1 and T3 using the procedure reported above.

Quantitative Real-Time PCR. The abundance of microorganisms identified in W001 was evaluated through quantitative real-time PCR (qPCR) in fecal sample F01. The DNA of F01 was extracted and diluted at a concentration of 10 ng. The presence of *Bradyrhizobium* spp., *Sphingobium* spp., and *Curvibacter* spp. DNA was evaluated using q-PCR with primer pair (5′-TGCGGTCACTCATCTTAGCT-3′/5′-GAGAACGCACGATCACCTTC-3′), (5′-CTGAACTGTTCGATCGGCTG-3′/5′-GCCATCGACCTCCTTATCCA-3′), and (5′-AGACCAGCTACAGATCGTCG-3′/5′-TACACATGCAAGTCGAACGG-3′), respectively. In detail, qPCR was performed using SoFast EvaGreen Supermix (Bio-Rad) on a CFX96 system (BioRad, CA, United States) following previously described protocols (Milani *et al.*, 2015). Each PCR reaction mix contained the following: 12.5 µl 2x SYBR SuperMix Green (BioRad, CA, United States), 5 µl of DNA at a concentration of 10 ng/µl, each of the forward and reverse primers at 0.5 µM, and nuclease-free water was added to obtain a final volume of 20 µl.

Statistical analysis. Bacterial abundance at the species level was validated by ANOVA analysis. Furthermore, PERMANOVA analysis was performed using 1000 permutations to estimate p-values of differences among infant samples in PCoA analyses. Statistical analyses were performed by using OriginPro graphing and analysis 2021.

Data availability

Shotgun metagenomics data are accessible through SRA study accession number PRJNA806724.

Reference

Bouchard, M.F., Sauvé, S., Barbeau, B., Legrand, M., Brodeur, M.È., Bouffard, T., et al. (2011) Intellectual impairment in school-age children exposed to manganese from drinking water. *Environ Health Perspect* **119**: 138–143.

Bowyer, R.C.E., Schillereff, D.N., Jackson, M.A., Le Roy, C., Wells, P.M., Spector, T.D., and Steves, C.J. (2020) Associations between UK tap water and gut microbiota composition suggest the gut microbiome as a potential mediator of health differences linked to water quality. *Sci Total Environ* **739**:.

Brumfield, K.D., Hasan, N.A., Leddy, M.B., Cotruvo, J.A., Rashed, S.M., Colwell, R.R., and Huq, A. (2020) A comparative analysis of drinking water employing metagenomics. *PLoS One* **15**:.

Busse, H.J., Denner, E.B.M., Buczolits, S., Salkinoja-Salonen, M., Bennasar, A., and Kämpfer, P. (2003) Sphingomonas aurantiaca sp. nov., Sphingomonas aerolata sp. nov. and Sphingomonas faeni sp. nov., air- and dustborne and Antarctic, orange-pigmented, psychrotolerant bacteria, and emended description of the genus Sphingomonas. *Int J Syst Evol Microbiol* **53**: 1253–1260.

Chan, S., Pullerits, K., Keucken, A., Persson, K.M., Paul, C.J., and Rådström, P. (2019) Bacterial release from pipe biofilm in a full-scale drinking water distribution system. *npj Biofilms Microbiomes* **5**:.

Chao, Y., Ma, L., Yang, Y., Ju, F., Zhang, X.X., Wu, W.M., and Zhang, T. (2013) Metagenomic analysis reveals significant changes of microbial compositions and protective functions during drinking water treatment. *Sci Rep* **3**:.

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

Dai, D., Rhoads, W.J., Edwards, M.A., and Pruden, A. (2018) Shotgun Metagenomics Reveals Taxonomic and Functional Shifts in Hot water microbiome due to temperature setting and stagnation. *Front Microbiol* **9**:.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., et al. (2021) Twelve years of SAMtools and BCFtools. *Gigascience* **10**:.

Dias, M.F., Reis, M.P., Acurcio, L.B., Carmo, A.O., Diamantino, C.F., Motta, A.M., et al. (2018) Changes in mouse gut bacterial community in response to different types of drinking water. *Water Res* **132**: 79–89.

Dimidi, E., Cox, S.R., Rossi, M., and Whelan, K. (2019) Fermented foods: Definitions and characteristics, impact on the gut microbiota and effects on gastrointestinal health and disease. *Nutrients* **11**:.

Dodd, M.C. (2012) Potential impacts of disinfection processes on elimination and deactivation of antibiotic resistance genes during water and wastewater treatment. *J Environ Monit* **14**: 1754–1771.

Eichler, S., Christen, R., Höltje, C., Westphal, P., Bötel, J., Brettar, I., et al. (2006) Composition and dynamics of bacterial communities of a drinking water supply system as assessed by RNA- and DNA-based 16S rRNA gene fingerprinting. *Appl Environ Microbiol* **72**: 1858–1872.

França, L., Lopéz-Lopéz, A., Rosselló-Móra, R., and da Costa, M.S. (2015) Microbial diversity and dynamics of a groundwater and a still bottled natural mineral water. *Environ Microbiol* **17**: 577–593.

Gulati, P. and Ghosh, M. (2017) Biofilm forming ability of Sphingomonas paucimobilis isolated from community drinking water systems on plumbing materials used in water distribution. *J Water Health* **15**: 942–954.

Hammes, F., Berney, M., Wang, Y., Vital, M., Köster, O., and Egli, T. (2008) Flow-cytometric total bacterial cell counts as a descriptive microbiological parameter for drinking water treatment processes. *Water Res* **42**: 269–277.

Hehemann, J.H., Correc, G., Barbeyron, T., Helbert, W., Czjzek, M., and Michel, G. (2010) Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**: 908–912.

Hillmann, B., Al-Ghalith, G.A., Shields-Cutler, R.R., Zhu, Q., Gohl, D.M., Beckman, K.B., et al. (2018) Evaluating the Information Content of Shallow Shotgun Metagenomics. *mSystems* **3**:.

Hong, P.Y., Hwang, C., Ling, F., Andersen, G.L., LeChevallier, M.W., and Liu, W.T. (2010) Pyrosequencing analysis of bacterial biofilm communities in water meters of a drinking water distribution system. *Appl Environ Microbiol* **76**: 5631–5635.

Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010) Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**:.

Kalmbach, S., Manz, W., Wecke, J., and Szewzyk, U. (1999) Aquabacterium gen. nov., with description of Aquabacterium citratiphilum sp. nov., Aquabacterium parvum sp. nov. and Aquabacterium commune sp. nov., three in situ dominant bacterial species from the Berlin drinking water system. *Int J Syst Bacteriol* **49**: 769–777.

Langdon, W.B. (2015) Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Min* **8**:.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Lorite, M.J., Estrella, M.J., Escaray, F.J., Sannazzaro, A., Videira E Castro, I.M., Monza, J., et al. (2018) The Rhizobia-Lotus symbioses: Deeply specific and widely diverse. *Front Microbiol* **9**:.

Loubet, P., Roux, P., Guérin-Schneider, L., and Bellon-Maurel, V. (2016) Life cycle assessment of forecasting scenarios for urban water management: A first implementation of the WaLA model on Paris suburban area. *Water Res* **90**: 128–140.

Loy, A., Beisker, W., and Meier, H. (2005) Diversity of bacteria growing in natural mineral water after bottling. *Appl Environ Microbiol* **71**: 3624–3632.

Lugli, G.A., Duranti, S., Milani, C., Mancabelli, L., Turroni, F., Alessandri, G., et al. (2020) Investigating bifidobacteria and human milk oligosaccharide composition of lactating mothers.

FEMS Microbiol Ecol **96**:.

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

Makki, K., Deehan, E.C., Walter, J., and Bäckhed, F. (2018) The Impact of Dietary Fiber on Gut Microbiota in Host Health and Disease. *Cell Host Microbe* **23**: 705–715.

McAlister, M.B., Kulakov, L.A., O'Hanlon, J.F., Larkin, M.J., and Ogden, K.L. (2002) Survival and nutritional requirements of three bacteria isolated from ultrapure water. *J Ind Microbiol Biotechnol* **29**: 75–82.

Michán, C., Blasco, J., and Alhama, J. (2021) High-throughput molecular analyses of microbiomes as a tool to monitor the wellbeing of aquatic environments. *Microb Biotechnol* **14**: 870–885.

Milani, C., Andrea Lugli, G., Duranti, S., Turroni, F., Mancabelli, L., Ferrario, C., et al. (2015) Bifidobacteria exhibit social behavior through carbohydrate resource sharing in the gut. *Sci Rep* **5**:.

Milani, C., Casey, E., Lugli, G.A., Moore, R., Kaczorowska, J., Feehily, C., et al. (2018) Tracing mother-infant transmission of bacteriophages by means of a novel analytical tool for shotgun metagenomic datasets: METAnnotatorX. *Microbiome* **6**:.

Milani, C., Duranti, S., Napoli, S., Alessandri, G., Mancabelli, L., Anzalone, R., et al. (2019) Colonization of the human gut by bovine bacteria present in Parmesan cheese. *Nat Commun* **10**:.

Milani, C., Lugli, G.A., Fontana, F., Mancabelli, L., Alessandri, G., Longhi, G., et al. (2021) METAnnotatorX2: a Comprehensive Tool for Deep and Shallow Metagenomic Data Set Analyses. *mSystems* **6**:.

Morohoshi, T., Sato, N., Iizumi, T., Tanaka, A., and Ikeda, T. (2017) Identification and characterization of a novel N-acyl-homoserine lactonase gene in Sphingomonas ursincola isolated from industrial cooling water systems. *J Biosci Bioeng* **123**: 569–575.

Pilot, L.R. (2012) Federal food, drug, and cosmetic act. *Pharm Law Desk Ref* 25–39.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., et al. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431–437.

Sala-Comorera, L., Caudet-Segarra, L., Galofré, B., Lucena, F., Blanch, A.R., and García-Aljaro, C. (2020) Unravelling the composition of tap and mineral water microbiota: Divergences between next-generation sequencing techniques and culture-based methods. *Int J Food Microbiol* **334**:.

Sheu, S.Y., Shiau, Y.W., Wei, Y.T., and Chen, W.M. (2013) Sphingobium fontiphilum sp. nov., isolated from a freshwater spring. *Int J Syst Evol Microbiol* **63**: 1906–1911.

Simonin, M., Voss, K.A., Hassett, B.A., Rocca, J.D., Wang, S.Y., Bier, R.L., et al. (2019) In search of microbial indicator taxa: shifts in stream bacterial communities along an urbanization gradient. *Environ Microbiol* **21**: 3653–3668.

Singh, H., Du, J., Yang, J.E., Yin, C.S., Kook, M.C., and Yi, T.H. (2015) Novosphingobium aquaticum sp. nov., isolated from lake water in Suwon, Republic of Korea. *Antonie van Leeuwenhoek, Int J Gen Mol Microbiol* **108**: 851–858.

Szewzyk, U., Szewzyk, R., Manz, W., and Schleifer, K.H. (2000) Microbiogical safety of drinking water. *Annu Rev Microbiol* **54**: 81–127.

Takeuchi, M., Hamana, K., and Hiraishi, A. (2001) Proposal of the genus Sphingomonas sensu stricto and three new genera, Sphingobium, Novosphingobium and Sphingopyxis, on the basis of phylogenetic and chemotaxonomic analyses. *Int J Syst Evol Microbiol* **51**: 1405–1417.

Vosloo, S., Huo, L., Anderson, C.L., Dai, Z., Sevillano, M., and Pinto, A. (2021) Evaluating de Novo Assembly and Binning Strategies for Time Series Drinking Water Metagenomes . *Microbiol Spectr* **9**:.

Wedemeyer, A., Kliemann, L., Srivastav, A., Schielke, C., Reusch, T.B., and Rosenstiel, P. (2017) An improved filtering algorithm for big read datasets and its application to single-cell assembly. *BMC Bioinformatics* **18**:.

Yoon, J.H., Lee, C.H., Yeo, S.H., and Oh, T.K. (2005) Sphingopyxis baekryungensis sp. nov., an orange-pigmented bacterium isolated from sea water of the Yellow Sea in Korea. *Int J Syst Evol Microbiol* **55**: 1223–1227.

Zhang, X., Xia, S., Ye, Y., and Wang, H. (2021) Opportunistic pathogens exhibit distinct growth dynamics in rainwater and tap water storage systems. *Water Res* **204**:.

Chapter 6

Multifactorial microvariability of the Italian raw milk cheese microbiota and implication for current regulatory scheme

Fontana F*, Longhi G*, Alessandri G, Lugli G.A, Mancabelli L, Tarracchini C, Viappiani A, Anzalone R, Ventura M, Turroni F, Milani C

The results of this chapter were published in mSystems, 2023 Jan 23; e0106822. doi: 10.1128/msystems.01068-22.

*These authors contributed equally.

Abstract

Raw milk cheese manufactory is strictly regulated in Europe by the Protected Designation of Origin (PDO) quality scheme, which protects indigenous food products based on geographical and biotechnological features.

This study encompassed the collection of 128 raw milk cheese samples across Italy to investigate the resident microbiome correlated to current PDO specification. Shotgun metagenomic approaches highlighted how the microbial communities are primarily linked to each cheesemaking site and consequently to the use of site-specific Natural Whey Cultures (NWCs), defined by a multifactorial set of local environmental factors rather than solely by cheese type or geographical origin that guide the current PDO specification.

Moreover, in-depth functional characterization of Cheese Community State Types (CCSTs) and comparative genomics efforts, including metagenomically assembled genomes (MAGs) of the dominant microbial taxa, revealed NWCs-related unique enzymatic profiles impacting the organoleptic features of the produced cheeses and availability of bioactive compounds to consumers, with putative health implications.

Thus, these results highlighted the need for a profound rethinking of the current PDO designation with a focus on the production site-specific microbial metabolism to understand and guarantee the organoleptic features of the final product recognized as PDO.

Importance

The Protected Designation of Origin (PDO) guarantees the traceability of food production processes, and that the production takes place in a well-defined restricted geographical area. Nevertheless, the organoleptic qualities of the same dairy products, i.e., cheeses under the same PDO denomination, differ between manufacturers. The final product's flavor and qualitative aspects can be related to the

resident microbial population, not considered by the PDO denomination. Here, we analyzed a complete set of different Italian cheeses produced from raw milk through shotgun sequencing in order to study the variability of the different microbial profiles resident in Italian PDO cheeses.

Furthermore, an in-depth functional analysis, along with a comparative genomic analysis, was performed in order to correlate the taxonomic information with the organoleptic properties of the final product. This analysis made it possible to highlight how the PDO denomination should be revisited to understand the effect that Natural Whey Cultures (NWCs), used in the traditional production of raw milk cheese and unique to each manufacturer, impacts on the organoleptic features of the final product.

For Supplementary Materials see the article published in *mSystems*.

Introduction

According to the European Food Safety Authority (EFSA), raw milk is defined as milk produced by farm animals, generally cows, sheep, goats and buffaloes, which has neither been heated above 40° C nor subjected to any other treatment having an equivalent effect on the milk-associated microbial community (1). Therefore, while direct consumption of raw milk can expose to microbiological hazards (2), the presence of endogenous living microorganisms is considered responsible for the complex and interesting organoleptic features of raw milk cheeses compared to those derived from pasteurized milk (3). In this context, raw milk cheesemaking is strictly regulated in Europe by the Protected Designation of Origin (PDO) product quality scheme, which links products to their geographic origins by ensuring production, processing, and preparation within a specific geographical area and following specific regulated procedures, employing expertise of local producers and raw materials from the geographical environment concerned.

In the case of raw milk cheeses, the key factor defining the resident microbial community is the use of back-slopping, which consists of using Natural Whey Cultures (NWCs) as bacterial starters instead of commercially available strains. NWCs consist of fermented milk harbouring a complex microbial community from the raw milk that is constantly added at each production cycle (4), similarly to the use and maintenance of sourdough in breadmaking. Due to its nature, NWCs are extremely variable in relation to each specific production site and modulated by local environmental factors (5).

In this context, the structural and physical-chemical modifications induced during fermentation of the milk matrix by the indigenous microbial communities originating from NWCs are the fundamental biochemical process responsible for the texture and other functional qualities of dairy products (6–9). Indeed, the organoleptic characteristics of fermented dairy products, such as texture, aroma and flavor depend on the profile of molecules released by the microbiome-driven chemical conversion of carbohydrates, lipids, fats, and proteins, typically contained in milk (10–16). Moreover, the profile of functional molecules released by the local microbiota during cheese ripening will be metabolized by the human cheese consumers, thus exerting relevant biological roles impacting systemically on the

human health and well-being. Yet, despite this marked relevance of the microbial metabolism in cheesemaking, the cheese microbiomes and their productions site-specific high variability are just marginally considered in the current PDO regulations.

Due to the importance of the cheese microbiota in cheesemaking, many efforts have been made to understand the taxonomic composition and functional role of the microbial communities found in Italian cheeses (17–19). Nevertheless, a comprehensive dissection of the genomic and functional biodiversity of the microbiota harbored by PDO raw milk cheeses produced across the Italian peninsula is still missing. For this reason, we sampled 128 PDO raw milk cheeses covering all the main Italian types of cheese products (20), whose microbial populations and corresponding metabolic potential have been assessed through shotgun metagenomics using both short- and long-read sequencing approaches.

Results and Discussion

Metagenomic characterization of the bacterial community of PDO Italian raw milk cheeses.

In the framework of this study, we collected up to five samples for each of the main PDO raw milk cheeses produced in Italy (Fig.1). These are artisanal raw milk cheeses produced following the PDO guidelines and employing a cheesemaking technique named back-slopping, in which a small portion of the previous batch of fermented milk is used to support the next fermentation step of raw milk without adding commercial bacterial starters (4). This approach consists of a pre-activated microbial starter, selected during multiple back-slopping cycles and thus historically unique to each cheesemaking site. Furthermore, as this microbial starter is kept in continuous growth thanks to the daily addition of fresh raw milk, it also adapts to local variables on a micro-geographical scale such as temperature and humidity levels, ultimately causing fluctuations in the final organoleptic features of the dairy product.

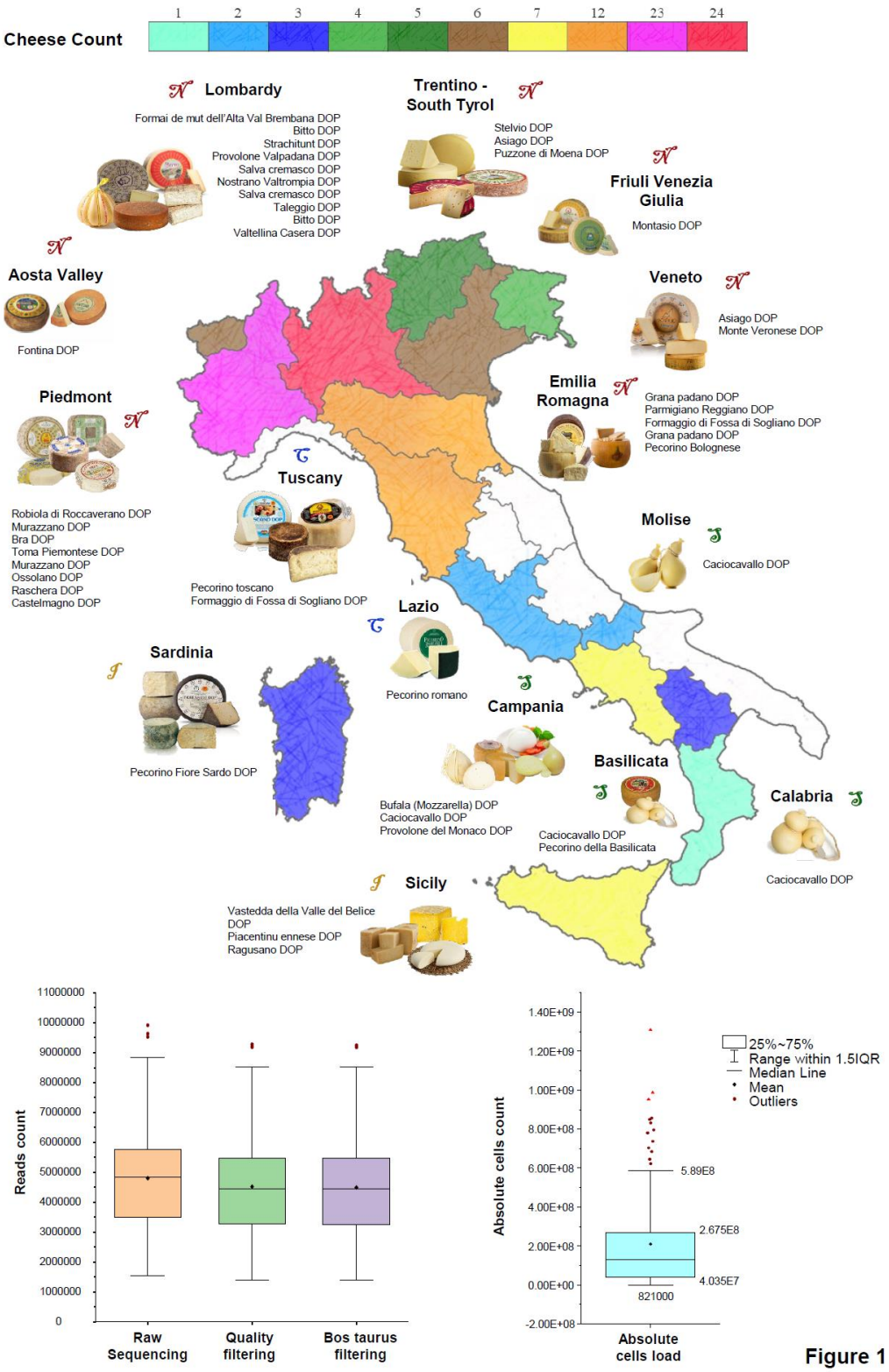


Fig.1. Geographical distribution of collected cheeses. In panel a) is reported a schematic representation of Italy, with regions colored according to the number of cheeses collected. For white-labelled regions, samples of cheeses have not been collected. Pictures of main cheeses from each region are reported. In panel b) is reported a Whisker plot representing the sequencing depth from raw to filtered reads, while in panel c) is reported a Whisker plot representing the absolute cells count distribution of each cheese.

Overall, we retrieved a total of 103 cheese samples corresponding to 32 PDO cheese types collected across the Italian peninsula, including multiple cheesemakers for cheese type (Fig.1) (Supplementary data). Furthermore, for comparison purposes, we also collected 25 samples of non-PDO cheeses, i.e., an (unpasteurized) raw-milk cheese type without PDO certification, which were manufactured with the artificial addition of selected microbial starters. Microbial DNA extracted from the collected samples was submitted to shotgun sequencing and raw reads were processed through the METAnnotatorX2 pipeline (21) in order to obtain species-level taxonomic profiles (Supplementary data) (Figure 1). Subsequently, a flow cytometry assay of the total bacterial load present in 0.2 gr of cheese was used to transform the relative abundance of each profiled microbial taxa into absolute abundance, i.e. estimation of species-specific cells load (Supplementary data) (Figure 1). Notably, no correlation was found between alpha diversity expressed as the number of observed species and PDO designation (Independent T-test p-value >0.05) (Supplementary data).

Multifactorial dissection of the species-level taxonomic composition across PDO and non-PDO Italian raw milk cheeses.

The species-level taxonomic composition of each cheese profile used in this study was explored to evaluate its variability across the Italian peninsula, considering both PDO and non-PDO cheeses. Intriguingly, prevalence analysis of bacterial species showed that 11 taxa could be found in at least 10% of the Italian PDO cheeses, corresponding to *Streptococcus thermophilus* (prevalence of 81.5 %), six *Lactobacillus* species (prevalence ranging from 12.6% % to 60.9 %), *Lactococcus lactis* (prevalence of 42.7%), *Lactiplantibacillus plantarum* (prevalence of 17.5 %), *Leuconostoc mesenteroides* (prevalence of 12.6 %) and *Bifidobacterium mongoliense* (prevalence of 11.6 %) (Supplementary data).

Notably, despite a core microbiota consisting of 11 highly prevalent species, visualization of the inter-sample's taxonomic diversity (beta-diversity) through a 2D Principal Coordinate Analysis (PCoA) revealed the absence of evident clustering of cheeses based on cheese type or regional localization

(Supplementary Fig.1). Nevertheless, validation through ANOSIM analysis revealed an R correlation of 12.8% (P <0.005) (Supplementary Fig.1) indicative that geographical region partially participate in defining the taxonomic composition. In-depth statistical investigation (detailed in the supplementary text) ultimately revealed that this result is due to the specific use of *Lactococcus lactis* as microbial starter in non-PDO cheeses from Tuscany, specifically Pecorino Toscano (Supplementary data). In contrast, no correlation between geographical region and cheese microbiota was found for PDO cheeses (Supplementary data).

To carry out a comprehensive and complete analysis, cheese matrix hardness was also evaluated as another high-relevant metadata, related directly to the ripening time, which may impact the cheese microbiota's taxonomic composition (22, 23). Therefore, each cheese sample was categorized as hard, semi-hard and soft cheese. This investigation highlighted that there is a correlation between matrix type and microbial composition (ANOSIM R 15.6%, P <0.001) (Supplementary Fig.3). Then, through a PCoA analysis, we noticed that most cheeses with hard matrices tend to cluster together. In contrast, semi-hard and soft cheeses did not show any particular clustering profile (Supplementary Fig.3). In detail, between the hard cheeses only two PDO types seem to cluster together, i.e. Parmigiano Reggiano and Grana Padano (Supplementary Fig.1 and Supplementary Fig.3). These two cheese types are hard and long-aged dairy products, which is a factor that leads to a decrease in the organic substrate initially present in the fresh, non-aged cheese matrix. As a result of this modification, a simplification of the resident microbiota occurs (average species richness of 6.5), which is reflected in the reduction of dispersion observed in the beta diversity analysis (Supplementary data) (Supplementary Fig.1 and Supplementary Fig. 3).

These observations highlight how the microbial particularities of the different cheese products with the same ripening stage are multifactorial and linked to the dairy site as a unique and comprehensive sum of each impacting factor while cheese aging will eventually induce a simplification of the microbial population. Nonetheless, further investigations are required to validate this approach, with particular focus on direct NWCs compositions and their seasonal composition stability.

Ecological investigation of co-occurrent microbial communities in Italian raw milk cheeses.

After evaluating the main metadata that could impact on the composition and stability of the cheese microbiota, the relative abundances of microbial profiles were normalized using the absolute cell load obtained from flow cytometry assays (Supplementary data) (Figure 1).

Then, to define microbial characteristics shared by different clusters of cheese samples, a Hierarchical Clustering Analysis (HCA) was performed based on their absolute abundance composition, leading to the definition of five High Prevalence Cheese Community State Types (HPCCSTs), i.e. high prevalence recurring microbial profiles, found in at least five among the 128 Italian raw milk cheeses collected in this study (Fig.2) (Supplementary data). The average bacterial load observed for the predicted HPCCSTs ranged from 6.14E+07 to 2.44E+08, (Supplementary data).

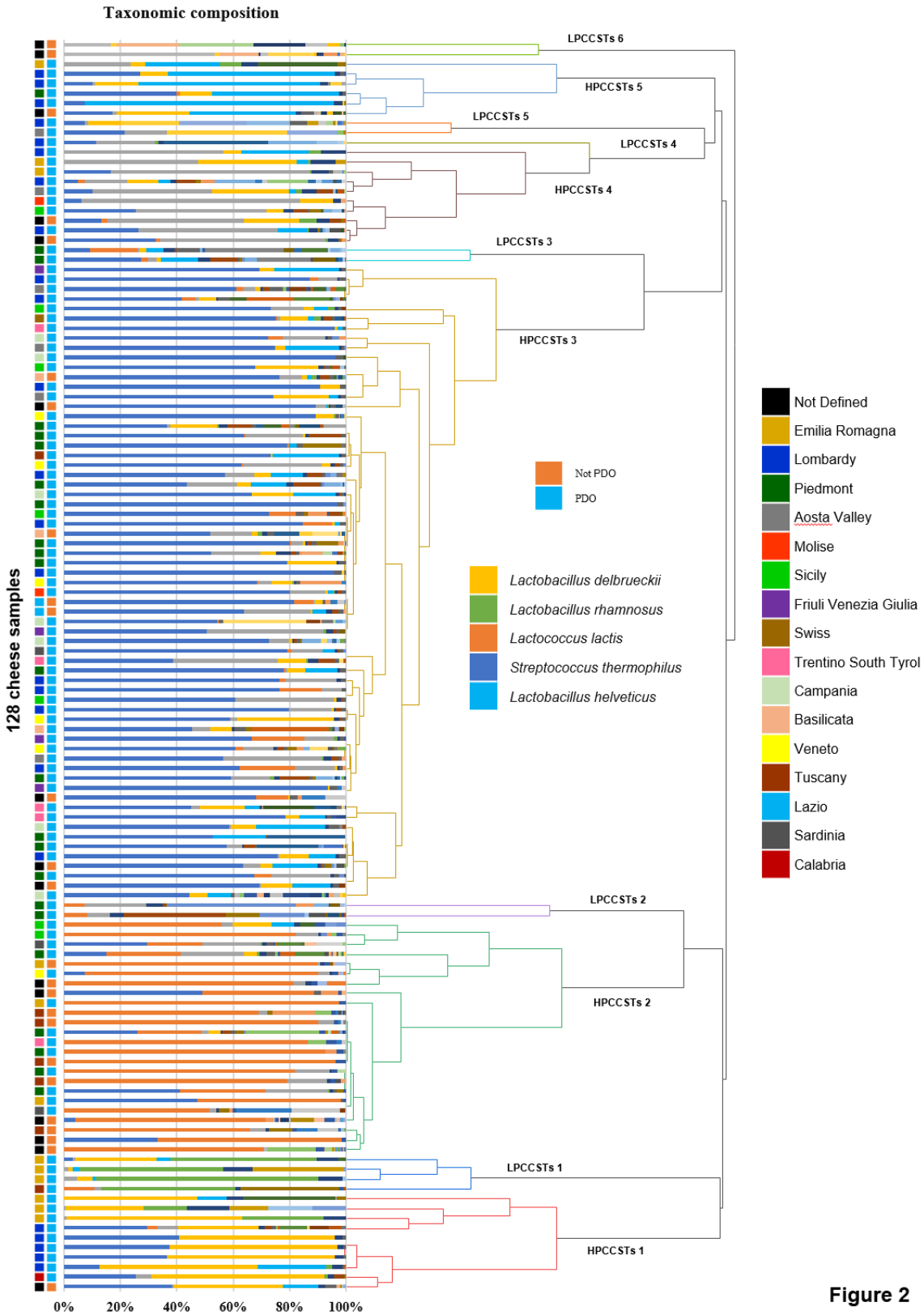
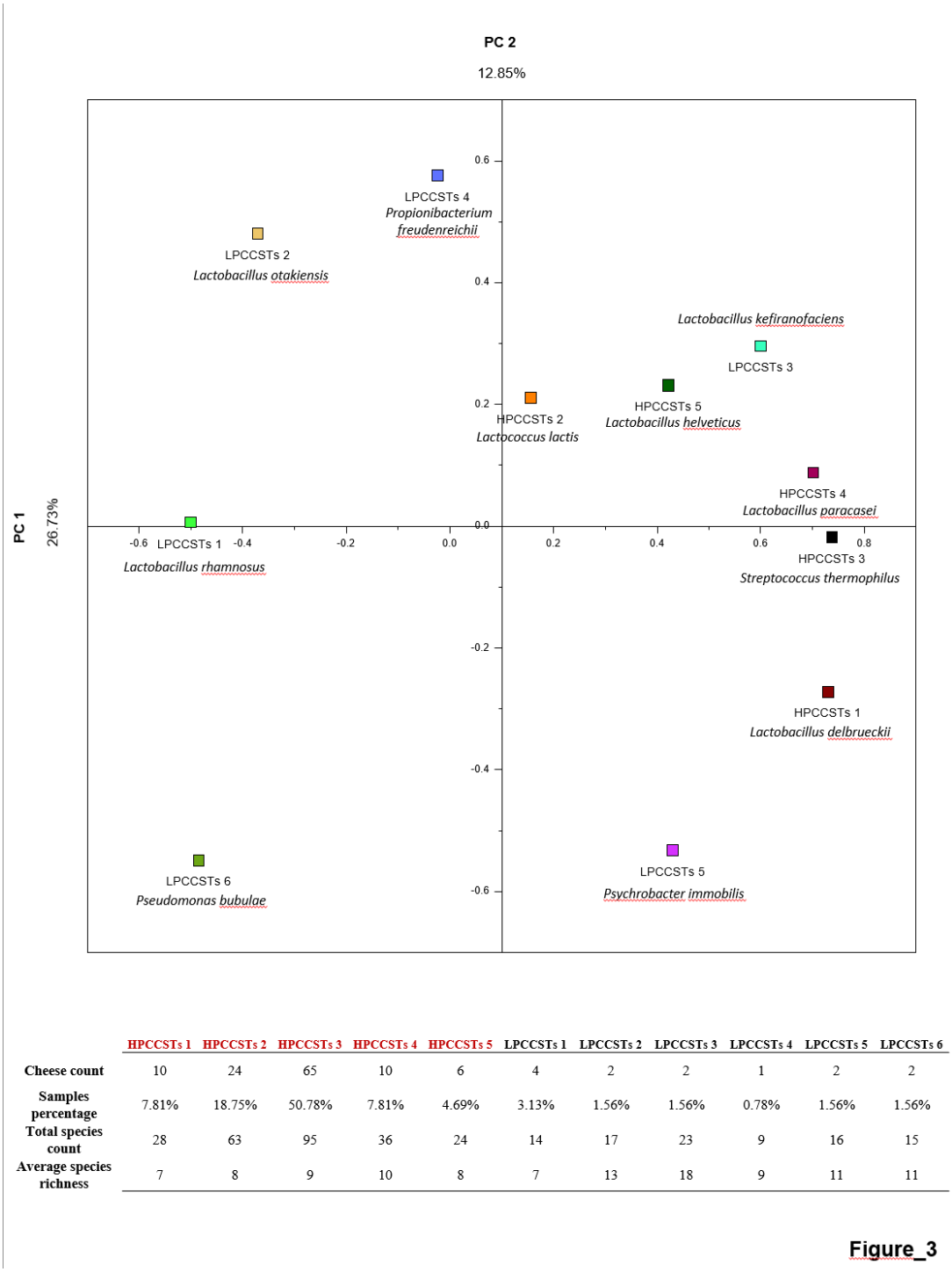


Fig. 2. HCL subdivision of all cheese samples. Graphic representation of HCL subdivision of cheese samples is reported, with branch colored based on HCA cluster. In addition, a stylized taxonomic profile of samples is shown along with PDO / non-PDO classification, geographical designation and legend of the main taxa are reported.

The five HPCCSTs are characterized by an average species richness ranging from seven to 10, with five species acting as (co)dominant by constituting on average >57 % of the HPCCSTs' microbial community along with the relevant participation of accessory taxa. In detail, *S. thermophilus* resulted dominant in HPCCST 3 and co-dominant in all the other four HPCCSTs, as expected by a thermophilic lactic acid bacteria (LAB) (24). Instead, *Lactobacillus* species *L. delbrueckii, L. paracasei* and *L. helveticus* as well as *Lactococcus lactis* act as dominant bacterial species in HPCCST 1, HPCCST 4, HPCCST 5 and HPCCST 2 respectively (Fig.2) (Fig.3) (Supplementary Fig.4) (Supplementary data).



	HPCCSTs 1	HPCCSTs 2	HPCCSTs 3	HPCCSTs 4	HPCCSTs 5	LPCCSTs 1	LPCCSTs 2	LPCCSTs 3	LPCCSTs 4	LPCCSTs 5	LPCCSTs 6
Cheese count	10	24	65	10	6	4	2	2	1	2	2
Samples percentage	7.81%	18.75%	50.78%	7.81%	4.69%	3.13%	1.56%	1.56%	0.78%	1.56%	1.56%
Total species count	28	63	95	36	24	14	17	23	9	16	15
Average species richness	7	8	9	10	8	7	13	18	9	11	11

Figure_3

Fig. 3. PCoA of CCSTs Bray Curtis dissimilarity matrix.

PCoA representation of beta diversity among the different CCSTs acts as a centroid for all the samples belonging to each CCST. Each CCST showed an average absolute composition based on the samples' absolute cell composition. Furthermore, the beta diversity score was based on a Bray-Curtis dissimilarity matrix to collapse the weight of each bacterial species into a single microbiological distance value to normalize the results and highlight the macro differences in microbial composition among the various CCSTs. Finally, near each CCSTs square point is also reported the predominant bacterial species for each CCST, as well as a summary of the main data regarding CCSTs species richness and sample count.

Furthermore, the HLC analysis also revealed six Low Prevalence CCSTs (LPCCSTs) supported each by less than five cheese samples (Fig.2) (Fig.3) (Supplementary Fig.4) (Supplementary data). In detail, LPCCSTs 5 and 6 represent clusters of contaminants that can be typically found in dairy production (Fig.2) (Fig.3) (Supplementary data) (25, 26).

As expected, evaluation of the distribution of non-PDO cheeses showed that they fall mainly in HPCCSTs 2 and 3 dominated by *L. lactis* and *S. thermophilus*, which are amongst the most common species exploited as artificial microbial starters in cheese manufacturing (27, 28) (Supplementary Fig.5) (Supplementary data). Subsequent statistical analyses were performed considering only PDO cheeses falling in the predicted CCSTs. Notably, we could not identify any clear correlations between cheese types or geographical origins and specific HPCCSTs, remarking that each production site has a major role in defining the cheese microbiota (Supplementary Fig.5). In addition, when the type of cheese matrix type (Soft, Semi-Hard and Hard) was correlated with the predicted HPCCSTs, it resulted that only Semi-Hard cheeses weakly and positively correlate (cor. 0.2037) with HPCCST 3 ($P < 0.05$) (Supplementary data).

These data confirm that cheese type-specific cheesemaking practices and cheese-related features like diary-matrix hardness have limited impact on the final microbial population harbored by the Italian raw milk cheeses collected. Instead, we propose that the micro-geographical uniqueness of each cheesemaking site over the cheese-type denomination represents the main driving force, with a putative key role of NWCs modulated by their unique local environmental factors (moisture, temperature etc.), along with the microbiota that naturally harbor in the local raw milk.

In the framework of this study, we also investigated the relationship between the bacterial species resident in PDO cheeses and the HCPPSTs through a bivariate correlation analysis that allowed the dissection of their ecological relationships (additional exhaustive discussion can be found in supplementary text).

Reconstruction of the metabolic potential of PDO Italian raw milk cheese's microbiota involved in developing cheese's organoleptic features.

After identifying the most common taxonomic profiles, also known as CCSTs, and how their species correlate, we evaluated how these different taxonomic clusters can organoleptically influence the final cheese product through their microbial metabolism. Thus, shotgun metagenomics data of PDO cheeses were submitted to functional metabolic profiling by METAnnotatorX2 to evaluate the commitment of each HPCCSTs toward a manually curated database of enzymatic reactions. This process allowed to reconstruct a functional profile covering a total of 1746 enzymatic reactions that showed >5% prevalence between the pool of 128 cheese samples analyzed. Since the data used are based on shotgun metagenomics with high-depth sequencing, this functional analysis was able to trace genes present in extremely low number of copies in the whole metagenome (<0.000002% in relative abundance). Then, following a Pearson correlation analysis, we extracted a subset of 48 statistically significant enzymatic reactions (29) that participate in the establishment of the cheese's organoleptic features and correlate with at least one of the HPCCSTs (27–32) (Supplementary data) (Fig.4). The selection of these 48 enzymatic reactions from the correlation pool was performed manually, exploiting what is reported in the recent literature (33–36) and selecting relevant enzymes along with products and by-products of organoleptic interest. In detail, selected enzymatic reactions refer to flavor enhancer molecules like acetaldehyde, ethanol, lactate and acetoin, other than technical agents like LPS-related enzymes (enhancer of texture in yogurt and other fermented dairy products) (Supplementary data). Additional information concerning the selected enzymes and their correlation score with the HPCCSTs are available in the Supplementary data).

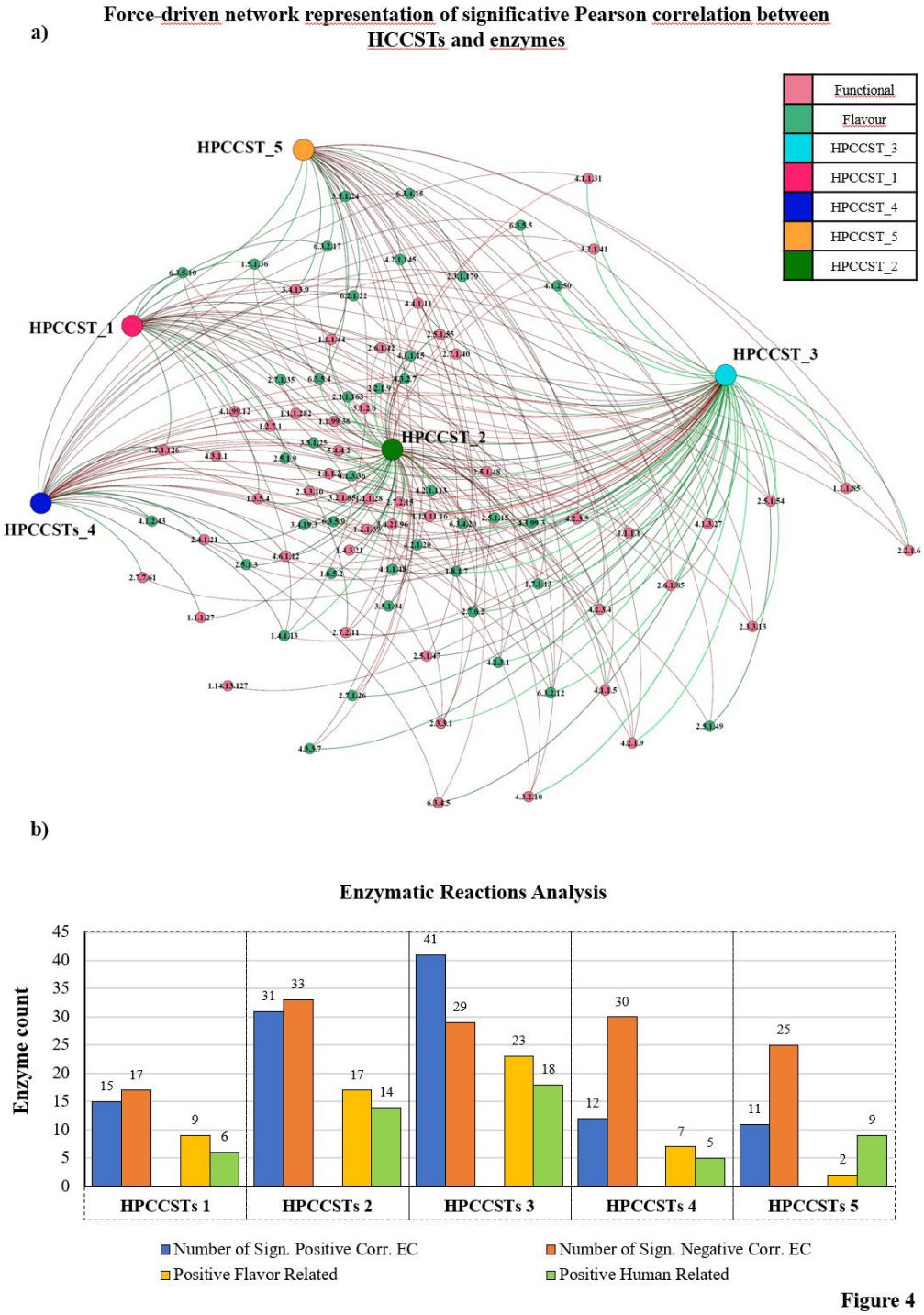


Figure 4

Fig. 4. Human and Flavor EC reports. In panel a) is reported a Network representation of correlation analysis based on a significative statistical relationship between the EC – numbers (enzymes) and HPCCSTs. Additionally, nodes were colored in order to separate flavor (green) and human health-related (pink) enzymes. In panel b) is reported a bar-plot graph showing correlations data regarding human health-supporting and flavor enzyme count and HPCCSTs. In detail, the blue bar represents the sum of all positive correlations between CCST and EC, the orange bar represents the sum of all negative correlations between CCST and EC, the yellow bar represents the sum of all positive correlations with EC numbers relating to the flavor enhancement and the green bar represents the sum of all positive correlations with EC numbers relating to human health-supporting functions (vitamin precursor etc.).

In detail, the number of positive correlations with enzymes inherent to organoleptically relevant flavors ranged from 2 (HPCCST_5) to 23 (HPCCST_3) (p.value < 0.001) (Fig.5). Notably, this result may represent the foundation of the differences in the organoleptic features observed for the same raw milk cheese type produced by different cheesemakers, as also suggested by the distribution of CCSTs across the collected types of cheese described above (Supplementary_Excell_File_3). Thus, emphasizing the key role in organoleptic features development exerted by specific microbial consortia. Specifically, once the microbiological profile has been categorized into one of the HPCCSTs categories, it is possible to trace a specific and expected metabolic potential in the final product, thus increasing our understanding of the possible organoleptic and health implications. Nonetheless, this needs to be confirmed through future RNA profiling and metabolomics studies regarding the actual expression of these 48 enzymes.

Subsequently, the average relative abundance of functional enzyme-encoding reads for each HPCCSTs analyzed was normalized using the absolute cell load obtained from flow cytometry data (Supplementary data) (Figure 1). This normalization of the functional profiles for the average bacterial load evidenced that the differences in average bacterial load observed for the predicted HPCCSTs (ranging from 6.14E+07 to 2.44E+08) may markedly impact their resulting metabolic activity (Supplementary data).

These data remark that the metabolic potential of the resident microbial population is probably linked to the manufacture-specific uniqueness (NWC and other environmental factors) (Supplementary data). Altogether, these results strengthen the notion that dissection of CCSTs composition and metabolic potential, coupled with bacterial load assessment, is a valuable target for food fingerprinting aimed at PDO cheese overall enhancement of the organoleptic and health-related features.

Predicted metabolites of raw milk cheese microbiota with potential impact on human physiology.

Recently, it has been demonstrated that the microbial community harbored by raw milk cheeses can colonize the gut of human consumers, where it can persist for weeks, especially when supported by a diet rich in milk and its derivates (37). Moreover, lactic acid bacteria (LAB) can also accumulate important secondary metabolites into cheese products, making them a natural supplement of important fermentation by-products (28, 32). For this reason, functional profiling of the cheeses' microbiota was employed to perform an explorative analysis of how each HPCCSTs-related enzymes may impact consumers' health. So, a subset of 40 enzymatic reactions which showed statistically relevant correlation and that lead to the production of high-interest microbial metabolites (38) was extracted (Supplementary data) (Fig.4).

In detail, among the 40 enzymes, selected manually based on recent scientific literature, there are enzymes participating in pathways that can lead to the production of vitamins or their precursors, such as the folate pathway (EC 2.5.1.15, related to Vitamin B9), the menaquinone-biosynthesis pathways (EC 2.1.1.163, related to Vitamin K2), flavin (EC 1.5.1.36, related to Vitamin B2) and a precursor of vitamin B12, adenosylcobyrate (EC 6.3.5.10) (41–43). Furthermore, there are other important molecules with putative functional effects on human health, such as molecules capable of reducing oxidative stress (EC 1.8.1.7, related to glutathione) (44–46) and molecules that can participate in the production of GABA (4-aminobutanoate and L-glutamate) (47, 48). Overall, the screening for enzymatic reactions encoded by the predicted HPCCSTs revealed a unique and significative correlation with enzymatic reaction patterns that support the role of raw milk cheeses as functional foods with a range of impacts on consumer health (Supplementary data).

These data support the drafting of future studies involving additional omics techniques, e.g. metabolomics, that will be pivotal in order to detailing the long-term impact of raw milk cheeses consumption on human health.

Genomic variability of the raw cheese microbiota across the Italian peninsula.

A comparative genomics analysis was performed to investigate further the genetic microbiome variability that characterizes each PDO cheese and their relationships with the geographical origin and cheese type. In addition, our analyses included metagenomically reconstructed genomes (MAGs). In detail, long reads sequencing was performed for 29 PDO and 10 non-PDO raw milk cheese samples collected across Italy. These cheeses were selected to cover the entire Italian peninsula, prioritizing selecting those cheeses with low species richness to allow efficient metagenomic assembly. Then, long reads were coupled with short reads metagenomics data to perform hybrid metagenomics assemblies that led to the reconstruction of draft genomes of the six most prevalent species profiled in raw milk cheeses (Supplementary Fig.6). Notably, 71 genomes were selected as they fullfill the average quality standards, i.e. showed > 90% of averaged completeness, with < 1% contamination and with > 94% of average ANI score respect to the species type strain. Thus, corresponding to a number of genomes ranging from 4 to 26 per species that were employed for comparative genomics analyses and pangenomes prediction (Supplementary data) (Supplementary Fig.6).

More than 10 genomes were retrieved from three species out of the six analyzed, i.e., *L. paracasei*, *L. delbrueckii* and *S. thermophilus*, and thus their unique gene content was analyzed (Supplementary data) (Fig.5).

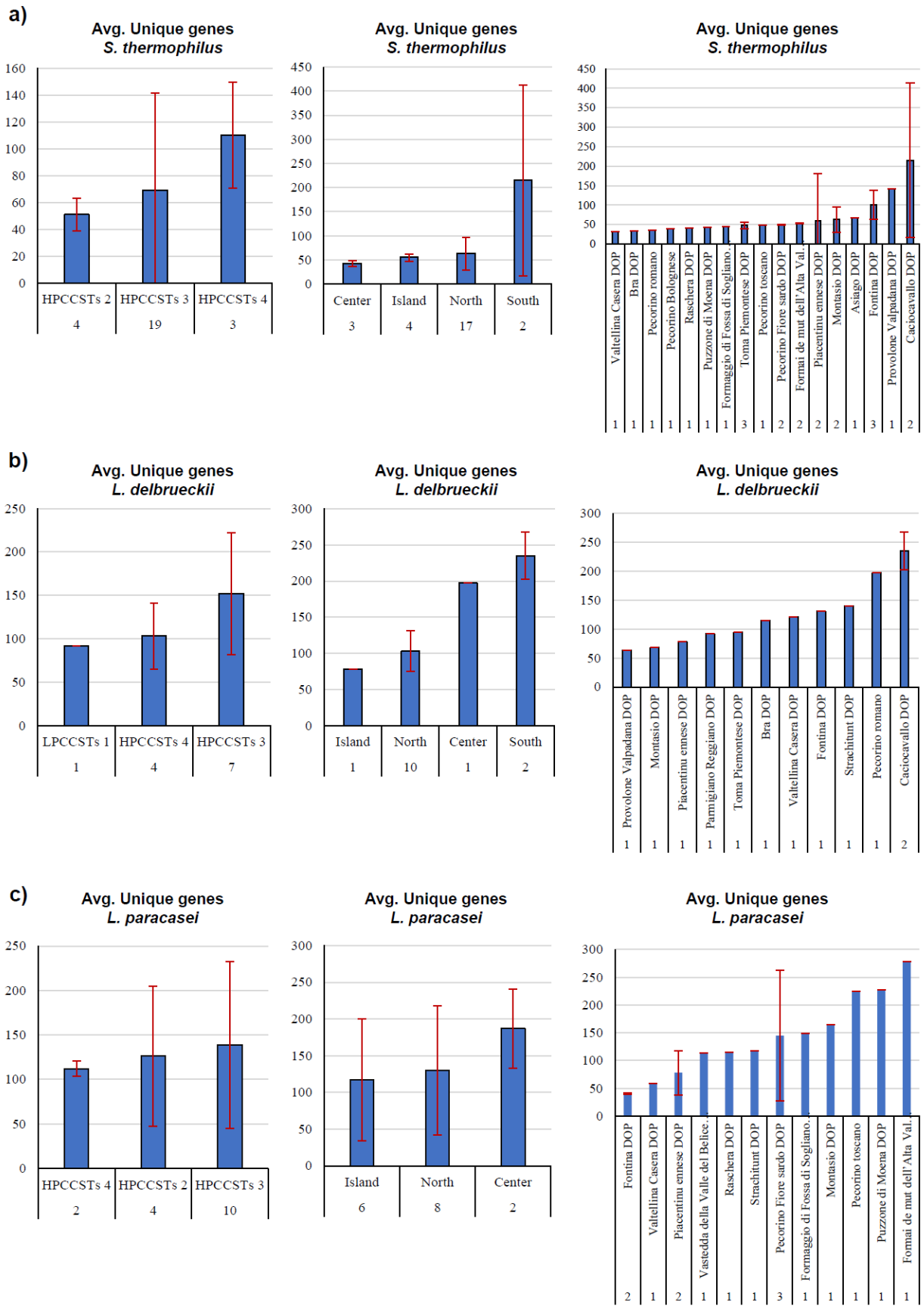

a)
Avg. Unique genes
S. thermophilus
HPCCSTs 2
HPCCSTs 3
HPCCSTs 4
4
19
3
Avg. Unique genes
S. thermophilus
Center
Island
North
South
3
4
17
2
Avg. Unique genes
S. thermophilus
Valtellina Casera DOP
Bra DOP
Pecorino romano
Pecorino Bolognese
Raschera DOP
Puzzone di Moena DOP
Formaggio di Fossa di Sogliano..
Toma Piemontese DOP
Pecorino toscano
Pecorino Fiore sardo DOP
Formai de mut dell'Alta Val..
Piacentinu ennese DOP
Montasio DOP
Asiago DOP
Fontina DOP
Provolone Valpadana DOP
Caciocavallo DOP
1 1 1 1 1 1 3 1 2 2 2 2 1 3 1 2
b)
Avg. Unique genes
L. delbrueckii
LPCCSTs 1
HPCCSTs 4
HPCCSTs 3
1
4
7
Avg. Unique genes
L. delbrueckii
Island
North
Center
South
1
10
1
2
Avg. Unique genes
L. delbrueckii
Provolone Valpadana DOP
Montasio DOP
Piacentinu ennese DOP
Parmigiano Reggiano DOP
Toma Piemontese DOP
Bra DOP
Valtellina Casera DOP
Fontina DOP
Strachitunt DOP
Pecorino romano
Caciocavallo DOP
1 1 1 1 1 1 1 1 1 1 2
c)
Avg. Unique genes
L. paracasei
HPCCSTs 4
HPCCSTs 2
HPCCSTs 3
2
4
10
Avg. Unique genes
L. paracasei
Island
North
Center
6
8
2
Avg. Unique genes
L. paracasei
Fontina DOP
Valtellina Casera DOP
Piacentinu ennese DOP
Vastedda della Valle del Belice..
Raschera DOP
Strachitunt DOP
Pecorino Fiore sardo DOP
Formaggio di Fossa di Sogliano..
Montasio DOP
Pecorino toscano
Puzzone di Moena DOP
Formai de mut dell'Alta Val..
2 1 2 1 1 1 3 1 1 1 1 1

Figure 5

Fig.5. Comparative genomics analysis on unique genes content and metadata subdivision. In panel a) are depicted three panels, showing the average unique genes content between *S. thermophilus* strains inside HPCCSTs clusters (first panel), between macro geographical area (second panel) and between cheese types (third panel), with the st.dev. reported when possible. In panel b) are reported three panels showing the average unique genes content between *L. delbueckii* strains inside HPCCSTs clusters (first panel), between macro geographical area (second panel) and between cheese types (third panel), with the st.dev. reported when possible. In panel c) are shown three panels showing the average unique genes content between *L. paracasei* strains inside HPCCSTs clusters (first panel), between macro geographical area (second panel) and between cheese types (third panel), with the st.dev. reported when possible.

Subsequently, PGAP pipeline (39) was used to obtain a Cluster of Orthologous Genes (COG) matrix, further processed in order to obtain the presence/absence of all retrieved genes. Then, the recovered matrix of genes presence/absence was used to profile the unique gene content of each genome (Supplementary data).

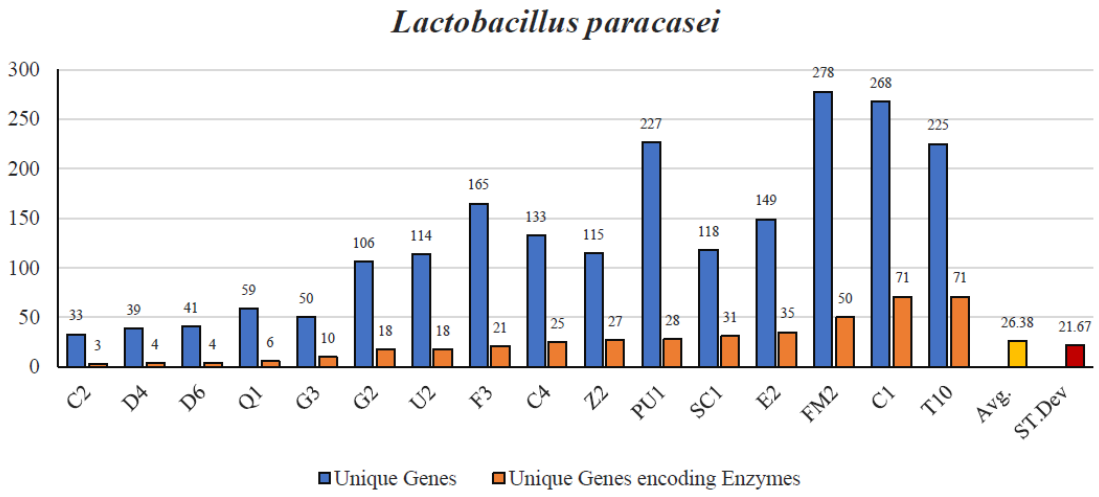
Additionally, based on the available metadata, Italian regions have been simplified to Islands, North, Central and South, and then crossed with the average content in unique genes (Fig. 5).

In detail, *L. paracasei* showed an average of unique genes of 117 (St.Dev of 83.3), 130 (St.Dev of 87.9) and 187 (St.Dev of 54.7) of strains assembled from cheese collected in Island, North and Center, respectively. Additionally, interpolation of comparative genomics results with other available metadata revealed that strains of the same species reconstructed from different cheeses type also showed high genetic variability, ranging from 40 to 278 unique genes content (Fig. 5). The same type of analysis was also performed for *S. thermophilus* and *L. delbrueckii*, displaying that the average content of unique genes showed a range from 31 to 215 for *S. thermophilus*, from 64 to 235 for *L. delbrueckii* and from 40 to 278 for *L. paracasei* (Fig. 5). However, a phylogenetic reconstruction based on the core genes content revealed close evolutionary relationships (Supplementary Fig.7).

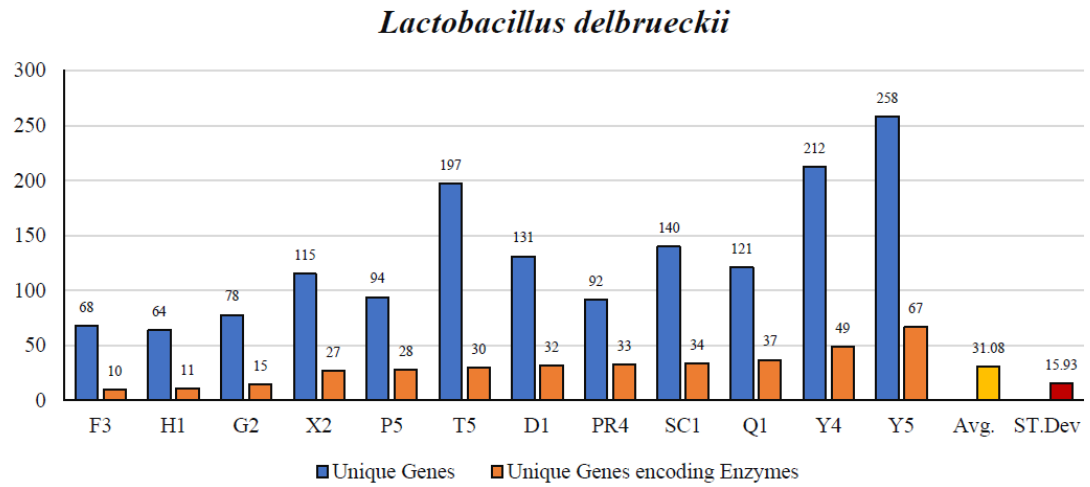
These results highlighted a marked genetic variability between different geographical areas and cheese types, supporting once again the role of cheesemaking site-specific NWCs adaptation to unique multifactorial environmental forces, including local raw milk microbiota, through cyclic back-slopping.

In this context, screening for enzymatic reactions (against MetaCyc enzyme database) showed that different strains of the same species also possess a unique enzymatic potential (Supplementary Fig.8). In detail, *L. paracasei*, *L. delbrueckii* and *S. thermophilus* strains showed an average of 26.4 (St.Dev of 21.6), 31.1 (St.Dev of 15.9) and 11.8 (St.Dev of 16.1) unique genes encoding for enzymes, respectively (Supplementary data) (Fig. 6).

a)


Lactobacillus paracasei
300
250
200
150
100
50
0
33 3
39 4
41 4
59 6
50 10
106 18
114 18
165 21
133 25
115 27
227 28
118 31
149 35
278 50
268 71
225 71
26.38
21.67
C2 D4 D6 Q1 G3 G2 U2 F3 C4 Z2 PU1 SC1 E2 FM2 C1 T10 Avg. ST.Dev
Unique Genes
Unique Genes encoding Enzymes

b)


Lactobacillus delbrueckii
300
250
200
150
100
50
0
68 10
64 11
78 15
115 27
94 28
197 30
131 32
92 33
140 34
121 37
212 49
258 67
31.08
15.93
F3 H1 G2 X2 P5 T5 D1 PR4 SC1 Q1 Y4 Y5 Avg. ST.Dev
Unique Genes
Unique Genes encoding Enzymes

c)

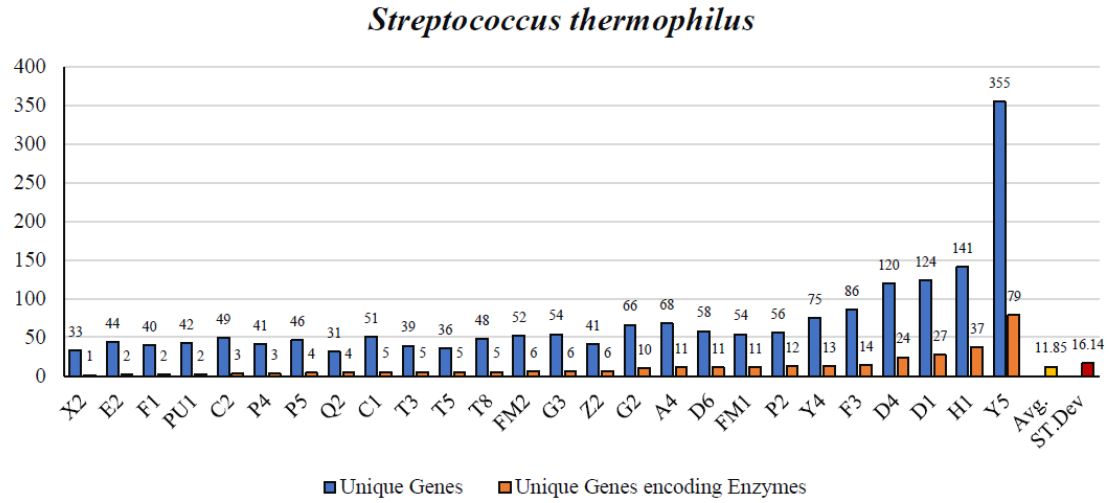

Streptococcus thermophilus
400
350
300
250
200
150
100
50
0
33 1
44 2
40 2
42 2
49 3
41 3
46 4
31 4
51 5
39 5
36 5
48 5
52 6
54 6
41 6
66 10
68 11
58 11
54 11
56 12
75 13
86 14
120 24
124 27
141 37
355 79
11.85
16.14
X2 E2 F1 PU1 C2 P4 P5 Q2 C1 T3 T5 T8 FM2 G3 Z2 G2 A4 D6 FM1 P2 Y4 F3 D4 D1 H1 Y5 Avg. ST.Dev
Unique Genes
Unique Genes encoding Enzymes

Figure 6

Fig.6. Unique genes content and enzymatic unique potential.
In panel a) is reported a bar plot showing the unique genes content (blue bar) for each different *L. paracasei* genomes tested, along with the unique genes encoding enzymes count (orange bar), the average unique genes encoding enzymes count (yellow bar) and the average St. Dev. (red bar). In panel b) is reported a bar plot showing the unique genes content (blue bar) for each different *L. delbrueckii* genomes tested, along with the unique genes encoding enzymes count (orange bar), the average unique genes encoding enzymes count (yellow bar) and the average St. Dev. (red bar). In panel c) is reported a bar plot showing the unique genes content (blue bar) for each different *S. thermophilus* genomes tested, along with the unique genes encoding enzymes count (orange bar), the average unique genes encoding enzymes count (yellow bar) and the average St. Dev. (red bar).

These strain-unique enzymatic features could be pivotal in the establishment of specific organoleptic features and in the development of bioactive compounds associated to each cheese producer.

Intriguingly, these data support the genetic uniqueness of the strains used to produce different types of PDO cheeses, which could be linked to the use of back-slopping techniques repeated for years as an alternative to commercial microbial starter strains. So, strains naturally present in the NWCs, and originating from the local raw milk microbiota, showed genetic adaptation to the complex set of environmental factor characterizing the production site (external factor as temperature, moisture, milk unique composition and bacterial competition in a semi-isolated system such as the diary factory production system), thus explaining the development of peculiar organoleptic features and potential metabolic profiles that differentiate the final products of each cheesemaker.

The results of this explorative analysis open the avenue of further intriguing future studies aimed at analyzing in detail the functionality of the here described genetic features that characterize different bacterial strains present in cheeses produced in different production sites and subject to different environmental factors.

Conclusions

The European PDO quality scheme protects regional raw milk cheese products by standardizing the cheesemaking process based on the know-how of local producers and ensuring that the manufacture is performed in a delimited geographical area using local ingredients. In this framework, increasing interest has recently grown regarding the resident cheese microbiota both for tracing and anti-counterfeit purposes and to disclose microbial communities' role in organoleptic features development and impact on human consumers.

To investigate these topics, we collected 128 raw milk cheeses across Italy for taxonomic and functional profiling of the resident microbiota. Results revealed how PDO cheeses of the same cheese type denomination but produced from different cheesemaking sites are characterized by unique microbial taxonomical, as well as microbial metabolically and genetic signatures that do not correlate only with their regional origin or cheese type. Instead, there is a vast set of multifactorial modulating factors behind the establishment of unique organoleptic features for each PDO cheese product tested, further linked to the unique composition of manufacturers-specific Natural Whey Cultures (NWCs) that can potentially be associated with the modulation of the final microbiological profiles. Factors that may impact the final taxonomical composition of the cheese products also include the raw milk microbiota used to maintain the NWC and additional environmental factors, such as moisture, temperature, milk composition and environmental contamination. Thus, the proposal of NWCs as a pivotal factor in the microbial imprinting on final cheese products will need to be confirmed with subsequent *ad hoc* studies.

Notably, these data contrast with the current PDO specification, which relies on the hypothesis of marked regional uniqueness for each specific cheese type denomination. In this way, while PDO certification can lead to the standardization of traditional production processes and guarantee their high-quality standard, it cannot ensure that the same cheese-type PDOs have the same organoleptic characteristics. In this regard, further studies should investigate the potential seasonality effects on

the finished product and microbial composition to gain a comprehensive overview of eventual seasonal confounding factors.

Altogether, these functional data underline that a better understanding of the metabolic potential of the microbial communities harbored by raw milk cheeses is pivotal not only for technological applications but also for obtaining dairy products with a high-value content of bioactive molecules that could influence the heath of the cheese-consumers.

Materials & Methods

Sample collection.

A total of 128 Italian cheese samples produced from raw milk were collected from different cheese maker encompassing large part of the diversity of Italian raw cheese production, considering the main cheese types, different producers, different geographical regions and both handmade and industrial productive processes. Between one and five samples belonging to different geographical places were collected for each type of Italian PDO raw milk cheese. More details regarding the variety of cheeses have been reported in Supplementary Table 1. By definition, each sample of cheese is not pasteurized and therefore is not subjected to any heat treatment in order to preserve the bacterial vitality. No precise information regarding temperature of acidification is available since every cheese maker may choose a specific one. Moreover, sample collection focused on cheese certified as PDO (Protected Designation of Origin), which must respect strict regulations specific for each cheese type that are aimed at preserving artisanal cheesemaking. During December 2019 and January 2020, almost 200 g of each cheese product were kept on ice and shipped to the laboratory under frozen conditions and vacuum packaged, after that they were preserved at –80°C, until they were processed.

Bacterial DNA extraction and Shotgun metagenomics sequencing.

Trying to avoid the rind, a fixed amount of 1 g of cheese belonging to the central portion was homogenized with 9 ml of phosphate buffered saline (PBS; pH 6.5). Subsequently, 1.5 mL of each resuspended cheese sample was subjected to bacterial DNA extraction using a DNeasy PowerFood

microbial kit according to the manufacturer's instructions (Qiagen, Germany). Then, each cheese sample's DNA concentration and purity was investigated by employing a Picodrop microtiter Spectrophotometer (Picodrop, Hinxton, UK). The extracted DNA was prepared using the Illumina Nextera XT DNA library preparation kit. Briefly, the DNA samples were enzymatically fragmented to 550–650 bp using a BioRuptor machine (Diagenode, Belgium), barcoded, and purified involving the Agencourt AMPure XP DNA purification beads (Beckman Coulter Genomics GmbH, Bernried, Germany). Then, samples were quantified using the fluorometric Qubit quantification system (Life Technologies, USA), loaded on a 2200 TapeStation instrument (Agilent Technologies, USA), and normalized to 4 nM. Sequencing was performed using an Illumina NextSeq 500 sequencer with NextSeq high output v2 kit chemicals (150 cycles) (Illumina Inc., San Diego, CA 92122, USA). All sequencing data were uploaded with BioProject PRJNA865096 and SRA study SRP389312.

Nanopore Sequencing and DNA processing

Approximately 1 μg of high molecular weight genomic DNA was used to prepare a sequencing library using the Ligation Sequencing Kit (SQK-LSK109) according to the manufacturer's instructions. For library cleanup, long fragment buffer (LFB) was used to retain DNA fragments. The sequencing library for DNA was prepared in conjunction with the Native barcoding genomic DNA (EXP-NBD104, EXP-NBD114), according to the manufacturer's instructions. Approximately 50 fmol of the prepared library was loaded onto the R9.4.1 flow cell. Sequencing was performed using the MinION Mk1B sequencing platform. Adaptive sequencing was applied using MinKNOW (21.10.6) software.

Metagenomics data processing

Taxonomic profiling of sequenced reads was performed with the METAnnotatorX2 bioinformatics platform (21, 40). In detail, the raw data in fastq format were submitted to quality filtering with removal of reads with an average quality <25. Subsequently, host DNA was removed by reads mapping to the *Bos taurus* genome. Finally, retained sequences were used as input to perform a MegaBLAST local alignment of reads to pre-processed database including available genomes of

eukaryotes (Fungi and Protists), bacteria, archaea, and viruses. Reads showing a nucleotide identity >94% to the genomes included in the database were classified at the species level, while if a lower percentage identity was detected, they were classified at the genus level as undefined species. These cut-offs are those generally employed for the ANI taxonomic assignment of genomes.

Functional profiling of sequenced reads was performed with the METAnnotatorX2 bioinformatics platform (21, 40) with an updated and manual curated enzymatic database, based on all available RefSeq genomes deposited on NCBI. DIAMOND software was used to assign Enzyme annotation with a MetaCyc updated database through the enzymatic code (EC) unique assignation.

Evaluation of bacterial cell density by flow cytometry

For total cell counts, 1 g of each cheese sample was resuspended and homogenized with PBS. Then, 1 mL of the initial homogenized cheese solution was 100,000 times diluted in physiological solution (PBS). Subsequently, 1 mL of the obtained bacterial cell suspension was stained with 1 µl of SYBR®Green I (ThermoFisher Scientific, USA) (1:100 dilution in dimethylsulfoxide; Sigma, Germany), vortex-mixed and incubated at 37 °C in the dark for at least 15 minutes before measurement. All count experiments were performed using an Attune NxT flow cytometry (ThermoFisher Scientific, Waltham, MA, USA) equipped with a blue laser set at 50 mW and tuned at an excitation wavelength of 488 nm. Multiparametric analyses were performed on both scattering signals, i.e., forward scatter (FSC) and side scatter (SSC), while SYBR Green I fluorescence was detected on the BL1 530/30 nm optical detector. Cell debris was excluded from acquisition analysis by setting a BL1 threshold. Furthermore, the gated fluorescence events were evaluated on the forward-sideways density plot to exclude remaining background events and to obtain an accurate microbial cell count, as previously described (41). All data were statistically analyzed with the Attune NxT flow cytometry software.

Statistics and Cluster analysis

HCL analysis was performed on OriginLabPro 2021b (42) with furthest neighbor and Pearson bivariate correlations, a type of analysis that highlight the linear relationships between pairs of continuous variables, ranging in strength and direction from -1 to 1 (43). Eigenvalues scores were retrieved from a Bray-Curtis dissimilarity matrix based on average relative abundance and/or absolute cells load normalized taxonomical profiles of samples, both obtained through the use of Rstudio (44) software. 3D and 2D PCoA representation of eigenvalues scores was made with OriginLabPro 2021b. PERMANOVA statistical analysis was performed on Rstudio (44) software. One Way ANOVA and Independent T-test were performed on SPSS software (55) with 1000 bootstrap. Pearson bivariate analysis was performed with Rstudios software and represented through a correlation Network made with Gephi software using Force Atlas 2 algorithm (56).

Comparative genomics analysis

Genome quality assessment was performed manually and through the use of checkM (57) software for completeness and contamination score, fastANI (58) software for the Average Nucleotide Identity between strains of the same species and sourmash (59) software for k-mer based genomes comparison. The pangenome and genes orthologous cluster analysis was performed through PGAP (49) software with --identity 0.5 and --coverage 0.8 as set up. DIAMOND (60) software was used for mapping unique genes protein sequences against a MetaCyc-derived EC database.

Data deposition

Raw sequences of shotgun data are accessible through SRA under BioProject number PRJNA865096.

Bibliography

1. Andreoletti O, Lau Baggesen D, Bolton D, Butaye P, Cook P, Davies R, Fernández Escámez PS, Griffin J, Hald T, Havelaar A, Koutsoumanis K, Lindqvist R, McLauchlin J, Nesbakken T, Prieto Maradona M, Ricci A, Ru G, Sanaa M, Simmons M, Sofos J, Barrucci F, Herman L, Hempen M, Stella P. 2015. Scientific Opinion on the public health risks related to the consumption of raw drinking milk. EFSA J 13:3940.
2. Verraes C, Vlaemynck G, Van Weyenberg S, De Zutter L, Daube G, Sindic M, Uyttendaele M, Herman L. 2015. A review of the microbiological hazards of dairy products made from raw milk. Int Dairy J 50:32–44.
3. Yoon Y, Lee S, Choi KH. 2016. Microbial benefits and risks of raw milk cheese. Food Control 63:201–215.
4. Olukotun GB, Salami SA, Okon IJ, Ahmadu JH, Ajibulu OO, Bello Z. 2021. Assessment of the Effects of Back Sloping on Some Starter Culture Strains and the Organoleptic Qualities of their Yoghurt Products. Asian Food Sci J 29–36.
5. Moser A, Schafroth K, Meile L, Egger L, Badertscher R, Irmler S. 2018. Population dynamics of Lactobacillus helveticus in Swiss Gruyère-type cheese manufactured with natural whey cultures. Front Microbiol 9:637.
6. Alegría Á, Szczesny P, Mayo B, Bardowski J, Kowalczyk M. 2012. Biodiversity in Oscypek, a traditional Polish Cheese, determined by culture-dependent and -independent approaches. Appl Environ Microbiol 78:1890–1898.
7. Delcenserie V, Taminiau B, Delhalle L, Nezer C, Doyen P, Crevecoeur S, Roussey D, Korsak N, Daube G. 2014. Microbiota characterization of a Belgian protected designation of origin cheese, Herve cheese, using metagenomic analysis. J Dairy Sci 97:6046–6056.
8. Giello M, La Storia A, Masucci F, Di Francia A, Ercolini D, Villani F. 2017. Dynamics of bacterial communities during manufacture and ripening of traditional Caciocavallo of Castelfranco cheese in relation to cows' feeding. Food Microbiol 63:170–177.

9. Shiby VK, Mishra HN. 2013. Fermented Milks and Milk Products as Functional Foods-A Review. Crit Rev Food Sci Nutr. Crit Rev Food Sci Nutr https://doi.org/10.1080/10408398.2010.547398.

10. Smit G, Smit BA, Engels WJM. 2005. Flavour formation by lactic acid bacteria and biochemical flavour profiling of cheese products. FEMS Microbiol Rev. Elsevier https://doi.org/10.1016/j.femsre.2005.04.002.

11. Grappin R, Beuvier E. 1997. Possible implications of milk pasteurization on the manufacture and sensory quality of ripened cheese. Int Dairy J. Elsevier https://doi.org/10.1016/S0958-6946(98)00006-5.

12. Carloni E, Petruzzelli A, Amagliani G, Brandi G, Caverni F, Mangili P, Tonucci F. 2016. Effect of farm characteristics and practices on hygienic quality of ovine raw milk used for artisan cheese production in central Italy. Anim Sci J 87:591–599.

13. Franciosi E, Settanni L, Cavazza A, Poznanski E. 2009. Presence of enterococci in raw cow's milk and "puzzone di moena" cheese. J Food Process Preserv 33:204–217.

14. Wouters JTM, Ayad EHE, Hugenholtz J, Smit G. 2002. Microbes from raw milk for fermented dairy products, p. 91–109. *In* International Dairy Journal.

15. De Angelis M, Corsetti A, Tosti N, Rossi J, Corbo MR, Gobbetti M. 2001. Characterization of Non-Starter Lactic Acid Bacteria from Italian Ewe Cheeses Based on Phenotypic, Genotypic, and Cell Wall Protein Analyses. Appl Environ Microbiol 67:2011–2020.

16. Lucchini R, Cardazzo B, Carraro L, Negrinotti M, Balzan S, Novelli E, Fasolato L, Fasoli F, Farina G. 2018. Contribution of natural milk culture to microbiota, safety and hygiene of raw milk cheese produced in alpine malga. Ital J Food Saf 7:55–61.

17. Cocconcelli PS, Fontana C, Bassi D, Gazzola S, Salvatore E. 2013. 25. Surface microbiota analysis of Italian cheeses 359–376.

18. Milani C, Fontana F, Alessandri G, Mancabelli L, Lugli GA, Longhi G, Anzalone R, Viappiani A, Duranti S, Turroni F, Ossiprandi MC, van Sinderen D, Ventura M. 2020.

Ecology of Lactobacilli Present in Italian Cheeses Produced from Raw Milk. Appl Environ Microbiol 86.

19. Pino A, Russo N, Solieri L, Sola L, Caggia C, Randazzo CL. 2022. Microbial Consortia Involved in Traditional Sicilian Sourdough: Characterization of Lactic Acid Bacteria and Yeast Populations. Microorganisms 10.
20. Cheeses PDO & PGI. https://www.dopitalianfood.com/en/brands-dop-italian-food/cheeses-pdo-pgi.html. Retrieved 2 August 2022.
21. 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.
22. Milani C, Fontana F, Alessandri G, Mancabelli L, Lugli GA, Longhi G, Anzalone R, Viappiani A, Duranti S, Turroni F, Ossiprandi MC, van Sinderen D, Ventura M. 2020. Ecology of lactobacilli present in italian cheeses produced from raw milk. Appl Environ Microbiol 86.
23. Mureşan CC, Marc RAV, Semeniuc CA, Socaci SA, Fărcaş A, Fracisc D, Pop CR, Rotar A, Dodan A, Mureşan V, Mureşan AE. 2021. Changes in physicochemical and microbiological properties, fatty acid and volatile compound profiles of apuseni cheese during ripening. Foods 10.
24. Montel M-C, Buchin S, Mallet A, Delbes-Paus C, Vuitton DA, Desmasures N, Berthier F. 2014. Traditional cheeses: Rich and diverse microbiota with associated benefits. Int J Food Microbiol 177:136–154.
25. Vidal AMC, Netto AS, Vaz ACN, Capodifóglio E, Gonçalves ACS, Rossi GAM, Figueiredo AS, Ruiz VLA. 2017. Pseudomonas spp.: contamination sources in bulk tanks of dairy farms. Pesqui Veterinária Bras 37:941–948.
26. Meng L, Zhang Y, Liu H, Zhao S, Wang J, Zheng N. 2017. Characterization of Pseudomonas spp. and associated proteolytic properties in raw milk stored at low temperatures. Front

Microbiol 8:2158.

27. Li W, Ren M, Duo L, Li J, Wang S, Sun Y, Li M, Ren W, Hou Q, Yu J, Sun Z, Sun T. 2020. Fermentation Characteristics of Lactococcus lactis subsp. lactis Isolated From Naturally Fermented Dairy Products and Screening of Potential Starter Isolates. Front Microbiol 11:1794.

28. Omae M, Maeyama Y, Nishimura T. 2008. Sensory Properties and Taste Compounds of Fermented Milk Produced by Lactococcus lactis and Streptococcus thermophilus. Food Sci Technol Res 14:183–189.

29. Enzyme Nomenclature. https://iubmb.qmul.ac.uk/enzyme/. Retrieved 6 April 2022.

30. Xu D, Ma M, Liu Y, Zhou T, Wang K, Deng Z, Hong K. 2015. PreQ0 base, an unusual metabolite with anti-cancer activity from Streptomyces qinglanensis 172205. Anticancer Agents Med Chem 15:285–290.

31. Atanasova J, Dalgalarrondo M, Iliev I, Moncheva P, Todorov SD, Ivanova I V. 2021. Formation of Free Amino Acids and Bioactive Peptides During the Ripening of Bulgarian White Brined Cheeses. Probiotics Antimicrob Proteins 13:261–272.

32. Murtaza MA, Ur-Rehman S, Anjum FM, Huma N, Hafiz I. 2014. Cheddar Cheese Ripening and Flavor Characterization: A Review. https://doi.org/101080/104083982011634531 54:1309–1321.

33. del Castillo-Lozano ML, Mansour S, Tâche R, Bonnarme P, Landaud S. 2008. The effect of cysteine on production of volatile sulphur compounds by cheese-ripening bacteria. Int J Food Microbiol 122:321–327.

34. Reis Lima MJ, Santos AO, Falcão S, Fontes L, Teixeira-Lemos E, Vilas-Boas M, Veloso ACA, Peres AM. 2019. Serra da Estrela cheese's free amino acids profiles by UPLC-DAD-MS/MS and their application for cheese origin assessment. Food Res Int 126:108729.

35. Balthazar CF, Guimarães JT, Silva R, Filho EGA, Brito ES, Pimentel TC, Rodrigues S, Esmerino EA, Silva MC, Raices RSL, Granato D, Duarte MCKH, Freitas MQ, Cruz AG.

2021. Effect of probiotic Minas Frescal cheese on the volatile compound and metabolic profiles assessed by nuclear magnetic resonance spectroscopy and chemometric tools. J Dairy Sci 104:5133–5140.

36. Smid EJ, Kleerebezem M. 2014. Production of Aroma Compounds in Lactic Fermentations. Annu Rev Food Sci Technol 5:313–326.

37. Wang J, Yang ZJ, Wang YD, Cao YP, Wang B, Liu Y. 2021. The key aroma compounds and sensory characteristics of commercial Cheddar Cheeses. J Dairy Sci https://doi.org/10.3168/jds.2020-19992.

38. Manzocchi E, Martin B, Bord C, Verdier-Metz I, Bouchon M, De Marchi M, Constant I, Giller K, Kreuzer M, Berard J, Musci M, Coppa M. 2021. Feeding cows with hay, silage, or fresh herbage on pasture or indoors affects sensory properties and chemical composition of milk and cheese. J Dairy Sci 104.

39. Milani C, Duranti S, Napoli S, Alessandri G, Mancabelli L, Anzalone R, Longhi G, Viappiani A, Mangifesta M, Lugli GA, Bernasconi S, Ossiprandi MC, van Sinderen D, Ventura M, Turroni F. 2019. Colonization of the human gut by bovine bacteria present in Parmesan cheese. Nat Commun 10:1–12.

40. Tunick MH, Van Hekken DL. 2015. Dairy Products and Health: Recent Insights. J Agric Food Chem 63:9381–9388.

41. Ryan-Harshman M, Aldoori W. 2008. Vitamin B12 and health. Can Fam Physician 54:536.

42. Daruwala R, Bhattacharyya DK, Kwon O, Meganathan R. 1997. Menaquinone (vitamin K2) biosynthesis: overexpression, purification, and characterization of a new isochorismate synthase from Escherichia coli. J Bacteriol 179:3133–3138.

43. Shams A. 2022. Folates: An Introduction. B-Complex Vitam - Sources, Intakes Nov Appl https://doi.org/10.5772/INTECHOPEN.102349.

44. Homma T, Fujii J. 2015. Application of Glutathione as Anti-Oxidative and Anti-Aging Drugs. Curr Drug Metab 16:560–571.

45. Baliou S, Adamaki M, Ioannou P, Pappa A, Panayiotidis MI, Spandidos DA, Christodoulou I, Kyriakopoulos AM, Zoumpourlis V. 2021. Protective role of taurine against oxidative stress (Review). Mol Med Rep 24.
46. Brosnan JT, Brosnan ME. 2006. The sulfur-containing amino acids: an overview. J Nutr 136.
47. Petroff OAC. 2002. GABA and glutamate in the human brain. Neuroscientist 8:562–573.
48. Brosnan JT, Brosnan ME. 2013. Glutamate: a truly functional amino acid. Amino Acids 45:413–418.
49. Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J. 2012. PGAP: Pan-genomes analysis pipeline. Bioinformatics 28:416–418.
50. 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.
51. Vandeputte D, Kathagen G, D'Hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, Wang J, Tito RY, De Commer L, Darzi Y, Vermeire S, Falony G, Raes J. 2017. Quantitative microbiome profiling links gut community variation to microbial load. Nature 551:507–511.
52. OriginLab. 2020. Origin 9.7.0.188: Scientific Data Analysis and Graphing Software. Orig Orig Introd . https://www.originlab.com/origin. Retrieved 14 July 2021.
53. Yeager K. LibGuides: SPSS Tutorials: Pearson Correlation.
54. RStudio | Open source & professional software for data science teams - RStudio. https://www.rstudio.com/. Retrieved 20 July 2022.
55. Software SPSS - Italia | IBM. https://www.ibm.com/it-it/analytics/spss-statistics-software. Retrieved 25 February 2021.
56. Gephi - The Open Graph Viz Platform. https://gephi.org/. Retrieved 25 February 2021.
57. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing

the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043.

58. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:1–8.

59. Brown CT, Irber L. 2016. sourmash: a library for MinHash sketching of DNA. J Open Source Softw 1:27.

60. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60.

Chapter 7

General Conclusion

Advances in the exploration of different microbial communities through the metagenomics approach

In the last years, great advances have been made in the knowledge of the composition and gene expression of the microbial component associated with various human body districts. Research in this field has rapidly risen thanks to the overcoming of the classical microbiology techniques limits through the new generation sequencing platforms that have allowed the simultaneous study of microbial communities at high resolution. Thanks to these new-generation techniques, the study of the intestinal microbiota has become a topic of current interest in the international scientific landscape following the discovery of its importance in multiple physiological processes. Indeed, it has been demonstrated that given its remarkable heritage of genes, the intestinal microbiota plays a crucial role in host health, acting as a barrier against pathogenic microorganisms, modulating the immune response and exercising central metabolic functions in the host organism, which for this reason must be considered as a complex system and therefore as a "superorganism". Perturbations in the composition and gene expression of the intestinal microbiota are associated with the risk of the onset of various disorders of the gastrointestinal tract, but they also seem to be involved in the onset and progression of important autoimmune diseases such as allergies, neurodegenerative and chronic neuro-inflammatory diseases. Therefore, the advent and continuous development of multi-omics technologies is providing revolutionary tools for studying the microbiota, highlighting many aspects inherent in its modulation and the multiple interactions with the external environment as well as food and pathogens. However, given its importance in human health, the need for good methodologies to obtain a reliable investigation of gut microbiota composition is constantly growing. A very weak point of metagenomics analysis is represented by the extraction of microbial DNA which should be highly representative of the real microbial biodiversity of the biological samples, especially when it comes to clinical samples for diagnostic purpose. Large amounts of human DNA in some biological specimens can interfere with the analysis of the bacterial content. Recently, significant scientific

literature focusing the interest on eukaryotic DNA depletion methods appears to be successful in host DNA removal in microbiome studies. Among host DNA depletion approaches, saponin-based protocols have been proposed as the golden standard. However, an important limitation to the use of saponin is related to the differential impact this reagent produces on bacterial lysis. Although this host DNA depletion method successfully reduced the amount of human DNA, we reported that saponin application drastically changed the microbial taxonomic profiles of different biological matrices generating artifacts. Specifically, we observed that saponin targets not only host cells but also Gram-negative bacterial cells, inducing a reduction in their abundance, probably due to their molecular structures making them more susceptible to lysis (Chapter 3).

As previously mentioned, the development of novel strategic approaches in NGS analysis has been useful in increasing knowledge about the microbial content of different communities. For this purpose, a potentially applied pipeline named Probiotic Identity Card (PIC) was proposed (Chapter 4). This new approach includes a combination of whole metagenome shotgun sequencing and flow cytometry analyses to characterize the bacterial contents in terms of presence as well as abundance and viability of probiotic products, revealing several inconsistencies and contaminations.

An exhaustive comprehension of the human intestinal microbiota also starts from the analysis of the multiple factors that can modulate its composition; for example, from the study of the microbial content of fermented foods like raw milk cheeses, which, through their organoleptic features and bioactive compounds can have implications on consumers' health. In this regard, we revealed that the microbiota sheltered by many Italian raw milk cheeses characterized by Protected Designation of Origin (PDO) is composed by microbial communities closely linked not only to the geographical region of origin or to the cheese type, but also to cheesemaking process, raw milk microbiota and environmental factors like temperature and humidity. From this study, we deduced that despite the production processes' standardization, cheese products maintain a high microbial biodiversity of origin. Finally, our interest was in evaluating the notion that water is not only crucial for our nutritional and physiological needs but may also be important as a vehicle of microorganisms to the

human gut and with this purpose, a metagenomic analysis of microbial communities residing in fresh potable water was performed (Chapter 5). Furthermore, we reported that a large part of the microorganisms present in water is represented by unknown bacteria that should require further investigations since through their genome reconstruction, they have also been identified as part of the human fecal microbiota. Therefore, this represents an interesting new scenario that needs to be investigated in great detail, as the sharing of microorganisms between tap water and the human gut microbiota can impact human health by modulating the gut microbiota.

References

1. Thursby E, Juge N. Introduction to the human gut microbiota. Biochem J. 2017;474(11):1823-36.
2. Jandhyala SM, Talukdar R, Subramanyam C, Vuyyuru H, Sasikala M, Nageshwar Reddy D. Role of the normal gut microbiota. World J Gastroenterol. 2015;21(29):8787-803.
3. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, et al. Metagenomic analysis of the human distal gut microbiome. Science. 2006;312(5778):1355-9.
4. Ventura M, Turroni F, Canchaya C, Vaughan EE, O'Toole PW, van Sinderen D. Microbial diversity in the human intestine and novel insights from metagenomics. Front Biosci (Landmark Ed). 2009;14(9):3214-21.
5. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol. 2016;14(8):e1002533.
6. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, et al. Diversity of the human intestinal microbial flora. Science. 2005;308(5728):1635-8.
7. Singh RK, Chang HW, Yan D, Lee KM, Ucmak D, Wong K, et al. Influence of diet on the gut microbiome and implications for human health. J Transl Med. 2017;15(1):73.
8. Alessandri G, Ossiprandi MC, MacSharry J, van Sinderen D, Ventura M. Bifidobacterial Dialogue With Its Human Host and Consequent Modulation of the Immune System. Front Immunol. 2019;10:2348.
9. Takiishi T, Fenero CIM, Camara NOS. Intestinal barrier and gut microbiota: Shaping our immune responses throughout life. Tissue Barriers. 2017;5(4):e1373208.
10. Pickard JM, Zeng MY, Caruso R, Nunez G. Gut microbiota: Role in pathogen colonization, immune responses, and inflammatory disease. Immunol Rev. 2017;279(1):70-89.
11. Topping DL, Clifton PM. Short-chain fatty acids and human colonic function: roles of resistant starch and nonstarch polysaccharides. Physiol Rev. 2001;81(3):1031-64.
12. Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI. Host-bacterial mutualism in the human intestine. Science. 2005;307(5717):1915-20.
13. Belkaid Y, Hand TW. Role of the microbiota in immunity and inflammation. Cell. 2014;157(1):121-41.
14. Lee YK, Mazmanian SK. Has the microbiota played a critical role in the evolution of the adaptive immune system? Science. 2010;330(6012):1768-73.
15. Liem NT, Agrawal V, Aison DS. Laparoscopic management of choledochal cyst in children: Lessons learnt from low-middle income countries. J Minim Access Surg. 2021;17(3):279-86.
16. Marcy Y, Ouverney C, Bik EM, Losekann T, Ivanova N, Martin HG, et al. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc Natl Acad Sci U S A. 2007;104(29):11889-94.
17. Alessandri G, Argentini C, Milani C, Turroni F, Cristina Ossiprandi M, van Sinderen D, et al. Catching a glimpse of the bacterial gut community of companion animals: a canine and feline perspective. Microb Biotechnol. 2020;13(6):1708-32.
18. Furrie E. A molecular revolution in the study of intestinal microflora. Gut. 2006;55(2):141-3.
19. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev. 2004;68(4):669-85.
20. Hamady M, Knight R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Res. 2009;19(7):1141-52.

21. Clarridge JE, 3rd. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev. 2004;17(4):840-62, table of contents.
22. Rintala A, Pietila S, Munukka E, Eerola E, Pursiheimo JP, Laiho A, et al. Gut Microbiota Analysis Results Are Highly Dependent on the 16S rRNA Gene Target Region, Whereas the Impact of DNA Extraction Is Minor. J Biomol Tech. 2017;28(1):19-30.
23. Heikema AP, Horst-Kreft D, Boers SA, Jansen R, Hiltemann SD, de Koning W, et al. Comparison of Illumina versus Nanopore 16S rRNA Gene Sequencing of the Human Nasal Microbiota. Genes (Basel). 2020;11(9).
24. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015;12(10):902-3.
25. Laudadio I, Fulci V, Palone F, Stronati L, Cucchiara S, Carissimi C. Quantitative Assessment of Shotgun Metagenomics and 16S rDNA Amplicon Sequencing in the Study of Human Gut Microbiome. OMICS. 2018;22(4):248-54.
26. Giannoukos G, Ciulla DM, Huang K, Haas BJ, Izard J, Levin JZ, et al. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. Genome Biol. 2012;13(3):R23.
27. Reck M, Tomasch J, Deng Z, Jarek M, Husemann P, Wagner-Dobler I, et al. Stool metatranscriptomics: A technical guideline for mRNA stabilisation and isolation. BMC Genomics. 2015;16(1):494.
28. Wilmes P, Bond PL. Microbial community proteomics: elucidating the catalysts and metabolic mechanisms that drive the Earth's biogeochemical cycles. Curr Opin Microbiol. 2009;12(3):310-7.
29. Ursell LK, Haiser HJ, Van Treuren W, Garg N, Reddivari L, Vanamala J, et al. The intestinal metabolome: an intersection between microbiota and host. Gastroenterology. 2014;146(6):1470-6.
30. Kolbert CP, Persing DH. Ribosomal DNA sequencing as a tool for identification of bacterial pathogens. Curr Opin Microbiol. 1999;2(3):299-305.
31. McCombie WR, McPherson JD, Mardis ER. Next-Generation Sequencing Technologies. Cold Spring Harb Perspect Med. 2019;9(11).
32. Takada H, Shimada T, Dey D, Quyyum MZ, Nakano M, Ishiguro A, et al. Differential Regulation of rRNA and tRNA Transcription from the rRNA-tRNA Composite Operon in Escherichia coli. PLoS One. 2016;11(12):e0163057.
33. Stewart FJ, Cavanaugh CM. Intragenomic variation and evolution of the internal transcribed spacer of the rRNA operon in bacteria. J Mol Evol. 2007;65(1):44-67.
34. Milani C, Lugli GA, Turroni F, Mancabelli L, Duranti S, Viappiani A, et al. Evaluation of bifidobacterial community composition in the human gut by means of a targeted amplicon sequencing (ITS) protocol. FEMS Microbiol Ecol. 2014;90(2):493-503.
35. Milani C, Duranti S, Mangifesta M, Lugli GA, Turroni F, Mancabelli L, et al. Phylotype-Level Profiling of Lactobacilli in Highly Complex Environments by Means of an Internal Transcribed Spacer-Based Metagenomic Approach. Appl Environ Microbiol. 2018;84(14).
36. Milani C, Alessandri G, Mangifesta M, Mancabelli L, Lugli GA, Fontana F, et al. Untangling Species-Level Composition of Complex Bacterial Communities through a Novel Metagenomic Approach. mSystems. 2020;5(4).
37. Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, et al. Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. Front Microbiol. 2016;7:459.
38. Franzosa EA, Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, et al. Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. Nat Rev Microbiol. 2015;13(6):360-72.

39. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35(9):833-44.
40. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A. 2011;108 Suppl 1(Suppl 1):4516-22.
41. Pinto AJ, Raskin L. PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. PLoS One. 2012;7(8):e43093.
42. Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. Appl Environ Microbiol. 2005;71(12):8966-9.
43. Suzuki MT, Giovannoni SJ. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl Environ Microbiol. 1996;62(2):625-30.
44. Qiu X, Wu L, Huang H, McDonel PE, Palumbo AV, Tiedje JM, et al. Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. Appl Environ Microbiol. 2001;67(2):880-7.
45. Wu JH, Hong PY, Liu WT. Quantitative effects of position and type of single mismatch on single base primer extension. J Microbiol Methods. 2009;77(3):267-75.
46. Thompson JR, Marcelino LA, Polz MF. Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. Nucleic Acids Res. 2002;30(9):2083-8.
47. Hugenholtz P, Huber T. Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. Int J Syst Evol Microbiol. 2003;53(Pt 1):289-93.
48. Polz MF, Cavanaugh CM. Bias in template-to-product ratios in multitemplate PCR. Appl Environ Microbiol. 1998;64(10):3724-30.
49. Krehenwinkel H, Wolf M, Lim JY, Rominger AJ, Simison WB, Gillespie RG. Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Sci Rep. 2017;7(1):17668.
50. Wojdacz TK, Hansen LL, Dobrovic A. A new approach to primer design for the control of PCR bias in methylation studies. BMC Res Notes. 2008;1:54.
51. Nichols RV, Vollmers C, Newsom LA, Wang Y, Heintzman PD, Leighton M, et al. Minimizing polymerase biases in metabarcoding. Mol Ecol Resour. 2018.
52. Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, et al. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nat Biotechnol. 2016;34(9):942-9.
53. Wylie KM, Truty RM, Sharpton TJ, Mihindukulasuriya KA, Zhou Y, Gao H, et al. Novel bacterial taxa in the human microbiome. PLoS One. 2012;7(6):e35294.
54. Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. Front Plant Sci. 2014;5:209.
55. Heravi FS, Zakrzewski M, Vickery K, Hu H. Host DNA depletion efficiency of microbiome DNA enrichment methods in infected tissue samples. J Microbiol Methods. 2020;170:105856.
56. Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome. 2018;6(1):42.
57. Ganda E, Beck KL, Haiminen N, Silverman JD, Kawas B, Cronk BD, et al. DNA Extraction and Host Depletion Methods Significantly Impact and Potentially Bias Bacterial Detection in a Biological Fluid. mSystems. 2021;6(3):e0061921.
58. Thoendel M, Jeraldo PR, Greenwood-Quaintance KE, Yao JZ, Chia N, Hanssen AD, et al. Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. J Microbiol Methods. 2016;127:141-5.

59. Hasan MR, Rawat A, Tang P, Jithesh PV, Thomas E, Tan R, et al. Depletion of Human DNA in Spiked Clinical Specimens for Improvement of Sensitivity of Pathogen Detection by Next-Generation Sequencing. J Clin Microbiol. 2016;54(4):919-27.
60. Wen Y, Xiao F, Wang C, Wang Z. The impact of different methods of DNA extraction on microbial community measures of BALF samples based on metagenomic data. Am J Transl Res. 2016;8(3):1412-25.
61. Cheli S, Napoli A, Clementi E, Montrasio C. DNA extraction from fresh and frozen plasma: an alternative for real-time PCR genotyping in pharmacogenetics. Mol Biol Rep. 2020;47(8):6451-5.
62. Israeli O, Makdasi E, Cohen-Gihon I, Zvi A, Lazar S, Shifman O, et al. A rapid high-throughput sequencing-based approach for the identification of unknown bacterial pathogens in whole blood. Future Sci OA. 2020;6(6):FSO476.
63. Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat Biotechnol. 2019;37(7):783-92.
64. Amar Y, Lagkouvardos I, Silva RL, Ishola OA, Foesel BU, Kublik S, et al. Pre-digest of unprotected DNA by Benzonase improves the representation of living skin bacteria and efficiently depletes host DNA. Microbiome. 2021;9(1):123.
65. Doughty EL, Sergeant MJ, Adetifa I, Antonio M, Pallen MJ. Culture-independent detection and characterisation of Mycobacterium tuberculosis and M. africanum in sputum samples using shotgun metagenomics on a benchtop sequencer. PeerJ. 2014;2:e585.
66. Votintseva AA, Bradley P, Pankhurst L, Del Ojo Elias C, Loose M, Nilgiriwala K, et al. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. J Clin Microbiol. 2017;55(5):1285-98.
67. Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, et al. Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. Genome Biol. 2016;17:41.
68. Morrow AL, Rangel JM. Human milk protection against infectious diarrhea: implications for prevention and clinical care. Semin Pediatr Infect Dis. 2004;15(4):221-8.
69. Moradi M, Kousheh SA, Almasi H, Alizadeh A, Guimaraes JT, Yilmaz N, et al. Postbiotics produced by lactic acid bacteria: The next frontier in food safety. Compr Rev Food Sci Food Saf. 2020;19(6):3390-415.
70. Milani C, Duranti S, Bottacini F, Casey E, Turroni F, Mahony J, 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).
71. Putignani L, Del Chierico F, Petrucca A, Vernocchi P, Dallapiccola B. The human gut microbiota: a dynamic interplay with the host from birth to senescence settled during childhood. Pediatr Res. 2014;76(1):2-10.
72. Jimenez E, Fernandez L, Marin ML, Martin R, Odriozola JM, Nueno-Palop C, et al. Isolation of commensal bacteria from umbilical cord blood of healthy neonates born by cesarean section. Curr Microbiol. 2005;51(4):270-4.
73. Aagaard K, Ma J, Antony KM, Ganu R, Petrosino J, Versalovic J. The placenta harbors a unique microbiome. Sci Transl Med. 2014;6(237):237ra65.
74. Munyaka PM, Khafipour E, Ghia JE. External influence of early childhood establishment of gut microbiota and subsequent health implications. Front Pediatr. 2014;2:109.
75. Avershina E, Storro O, Oien T, Johnsen R, Pope P, Rudi K. Major faecal microbiota shifts in composition and diversity with age in a geographically restricted cohort of mothers and their children. FEMS Microbiol Ecol. 2014;87(1):280-90.

76. Rautava S, Luoto R, Salminen S, Isolauri E. Microbial contact during pregnancy, intestinal colonization and human disease. Nat Rev Gastroenterol Hepatol. 2012;9(10):565-76.
77. Gronlund MM, Lehtonen OP, Eerola E, Kero P. Fecal microflora in healthy infants born by different methods of delivery: permanent changes in intestinal flora after cesarean delivery. J Pediatr Gastroenterol Nutr. 1999;28(1):19-25.
78. Penders J, Thijs C, Vink C, Stelma FF, Snijders B, Kummeling I, et al. Factors influencing the composition of the intestinal microbiota in early infancy. Pediatrics. 2006;118(2):511-21.
79. Mitsou EK, Kirtzalidou E, Oikonomou I, Liosis G, Kyriacou A. Fecal microflora of Greek healthy neonates. Anaerobe. 2008;14(2):94-101.
80. Hagg T, Gulati AK, Behzadian MA, Vahlsing HL, Varon S, Manthorpe M. Nerve growth factor promotes CNS cholinergic axonal regeneration into acellular peripheral nerve grafts. Exp Neurol. 1991;112(1):79-88.
81. Azad MB, Konya T, Maughan H, Guttman DS, Field CJ, Chari RS, et al. Gut microbiota of healthy Canadian infants: profiles by mode of delivery and infant diet at 4 months. CMAJ. 2013;185(5):385-94.
82. Turroni F, Milani C, Duranti S, Ferrario C, Lugli GA, Mancabelli L, et al. Bifidobacteria and the infant gut: an example of co-evolution and natural selection. Cell Mol Life Sci. 2018;75(1):103-18.
83. Heeney DD, Gareau MG, Marco ML. Intestinal Lactobacillus in health and disease, a driver or just along for the ride? Curr Opin Biotechnol. 2018;49:140-7.
84. Fallani M, Young D, Scott J, Norin E, Amarri S, Adam R, et al. Intestinal microbiota of 6-week-old infants across Europe: geographic influence beyond delivery mode, breast-feeding, and antibiotics. J Pediatr Gastroenterol Nutr. 2010;51(1):77-84.
85. Backhed 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. 2015;17(5):690-703.
86. Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, 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. 2010;107(26):11971-5.
87. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. Nature. 2012;489(7415):220-30.
88. 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. 2011;108 Suppl 1(Suppl 1):4578-85.
89. Spor A, Koren O, Ley R. Unravelling the effects of the environment and host genotype on the gut microbiome. Nat Rev Microbiol. 2011;9(4):279-90.
90. Costello EK, Stagaman K, Dethlefsen L, Bohannan BJ, Relman DA. The application of ecological theory toward an understanding of the human microbiome. Science. 2012;336(6086):1255-62.
91. Rodriguez JM, Murphy K, Stanton C, Ross RP, Kober OI, Juge N, et al. The composition of the gut microbiota throughout life, with an emphasis on early life. Microb Ecol Health Dis. 2015;26:26050.
92. Clemente JC, Ursell LK, Parfrey LW, Knight R. The impact of the gut microbiota on human health: an integrative view. Cell. 2012;148(6):1258-70.
93. Mancabelli L, Milani C, Lugli GA, Turroni F, Ferrario C, van Sinderen D, et al. Meta-analysis of the human gut microbiome from urbanized and pre-agricultural populations. Environ Microbiol. 2017;19(4):1379-90.
94. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature. 2011;473(7346):174-80.

95. Costea PI, Hildebrand F, Arumugam M, Backhed F, Blaser MJ, Bushman FD, et al. Enterotypes in the landscape of gut microbial community composition. Nat Microbiol. 2018;3(1):8-16.
96. Knights D, Ward TL, McKinlay CE, Miller H, Gonzalez A, McDonald D, et al. Rethinking "enterotypes". Cell Host Microbe. 2014;16(4):433-7.
97. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial community variation in human body habitats across space and time. Science. 2009;326(5960):1694-7.
98. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature. 2012;486(7402):222-7.
99. Martinez JE, Kahana DD, Ghuman S, Wilson HP, Wilson J, Kim SCJ, et al. Unhealthy Lifestyle and Gut Dysbiosis: A Better Understanding of the Effects of Poor Diet and Nicotine on the Intestinal Microbiome. Front Endocrinol (Lausanne). 2021;12:667066.
100. Craven M, Egan CE, Dowd SE, McDonough SP, Dogan B, Denkers EY, et al. Inflammation drives dysbiosis and bacterial invasion in murine models of ileal Crohn's disease. PLoS One. 2012;7(7):e41594.
101. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, et al. Richness of human gut microbiome correlates with metabolic markers. Nature. 2013;500(7464):541-6.
102. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun. 2017;8(1):1784.
103. Borenstein E, Kupiec M, Feldman MW, Ruppin E. Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proc Natl Acad Sci U S A. 2008;105(38):14482-7.
104. Freilich S, Kreimer A, Borenstein E, Yosef N, Sharan R, Gophna U, et al. Metabolic-network-driven analysis of bacterial ecological strategies. Genome Biol. 2009;10(6):R61.
105. Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat Microbiol. 2016;2:16180.
106. Nishino K, Nishida A, Inoue R, Kawada Y, Ohno M, Sakai S, et al. Analysis of endoscopic brush samples identified mucosa-associated dysbiosis in inflammatory bowel disease. J Gastroenterol. 2018;53(1):95-106.
107. Karlsson F, Tremaroli V, Nielsen J, Backhed F. Assessing the human gut microbiota in metabolic diseases. Diabetes. 2013;62(10):3341-9.
108. Vernocchi P, Del Chierico F, Putignani L. Gut Microbiota Profiling: Metabolomics Based Approach to Unravel Compounds Affecting Human Health. Front Microbiol. 2016;7:1144.
109. Li Q, Han Y, Dy ABC, Hagerman RJ. The Gut Microbiota and Autism Spectrum Disorders. Front Cell Neurosci. 2017;11:120.
110. Rutsch A, Kantsjo JB, Ronchi F. The Gut-Brain Axis: How Microbiota and Host Inflammasome Influence Brain Physiology and Pathology. Front Immunol. 2020;11:604179.
111. Meng C, Bai C, Brown TD, Hood LE, Tian Q. Human Gut Microbiota and Gastrointestinal Cancer. Genomics Proteomics Bioinformatics. 2018;16(1):33-49.
112. Halfvarson J, Brislawn CJ, Lamendella R, Vazquez-Baeza Y, Walters WA, Bramer LM, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol. 2017;2:17004.
113. Zaneveld JR, McMinds R, Vega Thurber R. Stress and stability: applying the Anna Karenina principle to animal microbiomes. Nat Microbiol. 2017;2:17121.
114. Mangin I, Bonnet R, Seksik P, Rigottier-Gois L, Sutren M, Bouhnik Y, et al. Molecular inventory of faecal microflora in patients with Crohn's disease. FEMS Microbiol Ecol. 2004;50(1):25-36.
115. Gulden E, Wong FS, Wen L. The gut microbiota and Type 1 Diabetes. Clin Immunol. 2015;159(2):143-53.

116. Turroni F, Peano C, Pass DA, Foroni E, Severgnini M, Claesson MJ, et al. Diversity of bifidobacteria within the infant gut microbiota. PLoS One. 2012;7(5):e36957.
117. Butel MJ, Suau A, Campeotto F, Magne F, Aires J, Ferraris L, et al. Conditions of bifidobacterial colonization in preterm infants: a prospective analysis. J Pediatr Gastroenterol Nutr. 2007;44(5):577-82.
118. Tojo R, Suarez A, Clemente MG, de los Reyes-Gavilan CG, Margolles A, Gueimonde M, et al. Intestinal microbiota in health and disease: role of bifidobacteria in gut homeostasis. World J Gastroenterol. 2014;20(41):15163-76.
119. 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. 2021;9(2):e0117621.
120. Walter J. Ecological role of lactobacilli in the gastrointestinal tract: implications for fundamental and biomedical research. Appl Environ Microbiol. 2008;74(16):4985-96.
121. Petrova MI, Lievens E, Malik S, Imholz N, Lebeer S. Lactobacillus species as biomarkers and agents that can promote various aspects of vaginal health. Front Physiol. 2015;6:81.
122. Gerritsen J, Smidt H, Rijkers GT, de Vos WM. Intestinal microbiota in human health and disease: the impact of probiotics. Genes Nutr. 2011;6(3):209-40.
123. Garrett WS, Gordon JI, Glimcher LH. Homeostasis and inflammation in the intestine. Cell. 2010;140(6):859-70.
124. Morgan XC, Kabakchiev B, Waldron L, Tyler AD, Tickle TL, Milgrom R, et al. Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. Genome Biol. 2015;16(1):67.
125. Reshef L, Kovacs A, Ofer A, Yahav L, Maharshak N, Keren N, et al. Pouch Inflammation Is Associated With a Decrease in Specific Bacterial Taxa. Gastroenterology. 2015;149(3):718-27.
126. Cheng Y, Ling Z, Li L. The Intestinal Microbiota and Colorectal Cancer. Front Immunol. 2020;11:615056.
127. Sears CL, Geis AL, Housseau F. Bacteroides fragilis subverts mucosal biology: from symbiont to colon carcinogenesis. J Clin Invest. 2014;124(10):4166-72.
128. Brennan CA, Garrett WS. Fusobacterium nucleatum - symbiont, opportunist and oncobacterium. Nat Rev Microbiol. 2019;17(3):156-66.
129. Gharbia SE, Shah HN, Lawson PA, Haapasalo M. Distribution and frequency of Fusobacterium nucleatum subspecies in the human oral cavity. Oral Microbiol Immunol. 1990;5(6):324-7.
130. Kostic AD, Chun E, Robertson L, Glickman JN, Gallini CA, Michaud M, et al. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. Cell Host Microbe. 2013;14(2):207-15.
131. Marchesi JR, Dutilh BE, Hall N, Peters WH, Roelofs R, Boleij A, et al. Towards the human colorectal cancer microbiome. PLoS One. 2011;6(5):e20447.
132. Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, et al. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Genome Res. 2012;22(2):299-306.
133. Relman DA. The human microbiome: ecosystem resilience and health. Nutr Rev. 2012;70 Suppl 1(Suppl 1):S2-9.
134. Gibson GR, Roberfroid MB. Dietary modulation of the human colonic microbiota: introducing the concept of prebiotics. J Nutr. 1995;125(6):1401-12.
135. Campbell JM, Fahey GC, Jr., Wolf BW. Selected indigestible oligosaccharides affect large bowel mass, cecal and fecal short-chain fatty acids, pH and microflora in rats. J Nutr. 1997;127(1):130-6.

136. Rycroft CE, Jones MR, Gibson GR, Rastall RA. A comparative in vitro evaluation of the fermentation properties of prebiotic oligosaccharides. J Appl Microbiol. 2001;91(5):878-87.
137. Davani-Davari D, Negahdaripour M, Karimzadeh I, Seifan M, Mohkam M, Masoumi SJ, et al. Prebiotics: Definition, Types, Sources, Mechanisms, and Clinical Applications. Foods. 2019;8(3).
138. Gibson GR, Probert HM, Loo JV, Rastall RA, Roberfroid MB. Dietary modulation of the human colonic microbiota: updating the concept of prebiotics. Nutr Res Rev. 2004;17(2):259-75.
139. van Loo J, Coussement P, de Leenheer L, Hoebregs H, Smits G. On the presence of inulin and oligofructose as natural ingredients in the western diet. Crit Rev Food Sci Nutr. 1995;35(6):525-52.
140. Costabile A, Kolida S, Klinder A, Gietl E, Bauerlein M, Frohberg C, et al. A double-blind, placebo-controlled, cross-over study to establish the bifidogenic effect of a very-long-chain inulin extracted from globe artichoke (Cynara scolymus) in healthy human subjects. Br J Nutr. 2010;104(7):1007-17.
141. Pokusaeva K, Fitzgerald GF, van Sinderen D. Carbohydrate metabolism in Bifidobacteria. Genes Nutr. 2011;6(3):285-306.
142. Turroni F, Ozcan E, Milani C, Mancabelli L, Viappiani A, van Sinderen D, et al. Glycan cross-feeding activities between bifidobacteria under in vitro conditions. Front Microbiol. 2015;6:1030.
143. Ryan SM, Fitzgerald GF, van Sinderen D. Transcriptional regulation and characterization of a novel beta-fructofuranosidase-encoding gene from Bifidobacterium breve UCC2003. Appl Environ Microbiol. 2005;71(7):3475-82.
144. Locascio RG, Ninonuevo MR, Kronewitter SR, Freeman SL, German JB, Lebrilla CB, et al. A versatile and scalable strategy for glycoprofiling bifidobacterial consumption of human milk oligosaccharides. Microb Biotechnol. 2009;2(3):333-42.
145. LoCascio RG, Desai P, Sela DA, Weimer B, Mills DA. Broad conservation of milk utilization genes in Bifidobacterium longum subsp. infantis as revealed by comparative genomic hybridization. Appl Environ Microbiol. 2010;76(22):7373-81.
146. Gallo A, Passaro G, Gasbarrini A, Landolfi R, Montalto M. Modulation of microbiota as treatment for intestinal inflammatory disorders: An uptodate. World J Gastroenterol. 2016;22(32):7186-202.
147. Lin AE, Autran CA, Szyszka A, Escajadillo T, Huang M, Godula K, et al. Human milk oligosaccharides inhibit growth of group B Streptococcus. J Biol Chem. 2017;292(27):11243-9.
148. Gonia S, Tuepker M, Heisel T, Autran C, Bode L, Gale CA. Human Milk Oligosaccharides Inhibit Candida albicans Invasion of Human Premature Intestinal Epithelial Cells. J Nutr. 2015;145(9):1992-8.
149. Guaraldi F, Salvatori G. Effect of breast and formula feeding on gut microbiota shaping in newborns. Front Cell Infect Microbiol. 2012;2:94.
150. Vandenplas Y, Zakharova I, Dmitrieva Y. Oligosaccharides in infant formula: more evidence to validate the role of prebiotics. Br J Nutr. 2015;113(9):1339-44.
151. Jinno S, Toshimitsu T, Nakamura Y, Kubota T, Igoshi Y, Ozawa N, et al. Maternal Prebiotic Ingestion Increased the Number of Fecal Bifidobacteria in Pregnant Women but Not in Their Neonates Aged One Month. Nutrients. 2017;9(3).
152. Rinne MM, Gueimonde M, Kalliomaki M, Hoppu U, Salminen SJ, Isolauri E. Similar bifidogenic effects of prebiotic-supplemented partially hydrolyzed infant formula and breastfeeding on infant gut microbiota. FEMS Immunol Med Microbiol. 2005;43(1):59-65.
153. Ben XM, Li J, Feng ZT, Shi SY, Lu YD, Chen R, et al. Low level of galacto-oligosaccharide in infant formula stimulates growth of intestinal Bifidobacteria and Lactobacilli. World J Gastroenterol. 2008;14(42):6564-8.

154. Sierra C, Bernal MJ, Blasco J, Martinez R, Dalmau J, Ortuno I, et al. Prebiotic effect during the first year of life in healthy infants fed formula containing GOS as the only prebiotic: a multicentre, randomised, double-blind and placebo-controlled trial. Eur J Nutr. 2015;54(1):89-99.
155. Kapiki A, Costalos C, Oikonomidou C, Triantafyllidou A, Loukatou E, Pertrohilou V. The effect of a fructo-oligosaccharide supplemented formula on gut flora of preterm infants. Early Hum Dev. 2007;83(5):335-9.
156. Reid G. Probiotics: definition, scope and mechanisms of action. Best Pract Res Clin Gastroenterol. 2016;30(1):17-25.
157. Hill C, Guarner F, Reid G, Gibson GR, Merenstein DJ, Pot B, et al. Expert consensus document. The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic. Nat Rev Gastroenterol Hepatol. 2014;11(8):506-14.
158. Ventura M, Turroni F, van Sinderen D. Probiogenomics as a tool to obtain genetic insights into adaptation of probiotic bacteria to the human gut. Bioeng Bugs. 2012;3(2):73-9.
159. Rajab S, Tabandeh F, Shahraky MK, Alahyaribeik S. The effect of lactobacillus cell size on its probiotic characteristics. Anaerobe. 2020;62:102103.
160. Martinez RC, Bedani R, Saad SM. Scientific evidence for health effects attributed to the consumption of probiotics and prebiotics: an update for current perspectives and future challenges. Br J Nutr. 2015;114(12):1993-2015.
161. Lugli GA, Mangifesta M, Mancabelli L, Milani C, Turroni F, Viappiani A, et al. Compositional assessment of bacterial communities in probiotic supplements by means of metagenomic techniques. Int J Food Microbiol. 2019;294:1-9.
162. Markowiak P, Slizewska K. Effects of Probiotics, Prebiotics, and Synbiotics on Human Health. Nutrients. 2017;9(9).
163. Wieers G, Belkhir L, Enaud R, Leclercq S, Philippart de Foy JM, Dequenne I, et al. How Probiotics Affect the Microbiota. Front Cell Infect Microbiol. 2019;9:454.
164. Gareau MG, Sherman PM, Walker WA. Probiotics and the gut microbiota in intestinal health and disease. Nat Rev Gastroenterol Hepatol. 2010;7(9):503-14.
165. Ng SC, Hart AL, Kamm MA, Stagg AJ, Knight SC. Mechanisms of action of probiotics: recent advances. Inflamm Bowel Dis. 2009;15(2):300-10.
166. Otte JM, Podolsky DK. Functional modulation of enterocytes by gram-positive and gram-negative microorganisms. Am J Physiol Gastrointest Liver Physiol. 2004;286(4):G613-26.
167. Champagne CP, Ross RP, Saarela M, Hansen KF, Charalampopoulos D. Recommendations for the viability assessment of probiotics as concentrated cultures and in food matrices. Int J Food Microbiol. 2011;149(3):185-93.
168. Kechagia M, Basoulis D, Konstantopoulou S, Dimitriadi D, Gyftopoulou K, Skarmoutsou N, et al. Health benefits of probiotics: a review. ISRN Nutr. 2013;2013:481651.
169. Lugli GA, Longhi G, Alessandri G, Mancabelli L, Tarracchini C, Fontana F, et al. The Probiotic Identity Card: A Novel "Probiogenomics" Approach to Investigate Probiotic Supplements. Front Microbiol. 2021;12:790881.
170. Chang CJ, Lin TL, Tsai YL, Wu TR, Lai WF, Lu CC, et al. Next generation probiotics in disease amelioration. J Food Drug Anal. 2019;27(3):615-22.
171. Zhai Q, Feng S, Arjan N, Chen W. A next generation probiotic, Akkermansia muciniphila. Crit Rev Food Sci Nutr. 2019;59(19):3227-36.
172. Derrien M, van Hylckama Vlieg JE. Fate, activity, and impact of ingested bacteria within the human gut microbiota. Trends Microbiol. 2015;23(6):354-66.

173. Ou J, Carbonero F, Zoetendal EG, DeLany JP, Wang M, Newton K, et al. Diet, microbiota, and microbial metabolites in colon cancer risk in rural Africans and African Americans. Am J Clin Nutr. 2013;98(1):111-20.
174. Park J, Cheon JH. Incidence and Prevalence of Inflammatory Bowel Disease across Asia. Yonsei Med J. 2021;62(2):99-108.
175. Carvalho-Wells AL, Helmolz K, Nodet C, Molzer C, Leonard C, McKevith B, et al. Determination of the in vivo prebiotic potential of a maize-based whole grain breakfast cereal: a human feeding study. Br J Nutr. 2010;104(9):1353-6.
176. Martinez I, Lattimer JM, Hubach KL, Case JA, Yang J, Weber CG, et al. Gut microbiome composition is linked to whole grain-induced immunological improvements. ISME J. 2013;7(2):269-80.
177. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2014;505(7484):559-63.
178. Milani C, Ferrario C, Turroni F, Duranti S, Mangifesta M, van Sinderen D, et al. The human gut microbiota and its interactive connections to diet. J Hum Nutr Diet. 2016;29(5):539-46.
179. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science. 2011;334(6052):105-8.
180. Statovci D, Aguilera M, MacSharry J, Melgar S. The Impact of Western Diet and Nutrients on the Microbiota and Immune Response at Mucosal Interfaces. Front Immunol. 2017;8:838.
181. Kabeerdoss J, Devi RS, Mary RR, Ramakrishna BS. Faecal microbiota composition in vegetarians: comparison with omnivores in a cohort of young women in southern India. Br J Nutr. 2012;108(6):953-7.
182. Matijasic BB, Obermajer T, Lipoglavsek L, Grabnar I, Avgustin G, Rogelj I. Association of dietary type with fecal microbiota in vegetarians and omnivores in Slovenia. Eur J Nutr. 2014;53(4):1051-64.
183. Granado-Lorencio F, Hernandez-Alvarez E. Functional Foods and Health Effects: A Nutritional Biochemistry Perspective. Curr Med Chem. 2016;23(26):2929-57.
184. Aslam H, Green J, Jacka FN, Collier F, Berk M, Pasco J, et al. Fermented foods, the gut and mental health: a mechanistic overview with implications for depression and anxiety. Nutr Neurosci. 2020;23(9):659-71.
185. Vitorino LC, Bessa LA. Technological Microbiology: Development and Applications. Front Microbiol. 2017;8:827.
186. Macori G, Cotter PD. Novel insights into the microbiology of fermented dairy foods. Curr Opin Biotechnol. 2018;49:172-8.
187. Chilton SN, Burton JP, Reid G. Inclusion of fermented foods in food guides around the world. Nutrients. 2015;7(1):390-404.
188. Mozaffarian D, Hao T, Rimm EB, Willett WC, Hu FB. Changes in diet and lifestyle and long-term weight gain in women and men. N Engl J Med. 2011;364(25):2392-404.
189. Chen M, Sun Q, Giovannucci E, Mozaffarian D, Manson JE, Willett WC, et al. Dairy consumption and risk of type 2 diabetes: 3 cohorts of US adults and an updated meta-analysis. BMC Med. 2014;12:215.
190. Eussen SJ, van Dongen MC, Wijckmans N, den Biggelaar L, Oude Elferink SJ, Singh-Povel CM, et al. Consumption of dairy foods in relation to impaired glucose metabolism and type 2 diabetes mellitus: the Maastricht Study. Br J Nutr. 2016;115(8):1453-61.
191. Soedamah-Muthu SS, Masset G, Verberne L, Geleijnse JM, Brunner EJ. Consumption of dairy products and associations with incident diabetes, CHD and mortality in the Whitehall II study. Br J Nutr. 2013;109(4):718-26.
192. Tapsell LC. Fermented dairy food and CVD risk. Br J Nutr. 2015;113 Suppl 2:S131-5.

193. Iwasa M, Aoi W, Mune K, Yamauchi H, Furuta K, Sasaki S, et al. Fermented milk improves glucose metabolism in exercise-induced muscle damage in young healthy men. Nutr J. 2013;12:83.
194. Voreades N, Kozil A, Weir TL. Diet and the development of the human intestinal microbiome. Front Microbiol. 2014;5:494.
195. Milani C, Duranti S, Napoli S, Alessandri G, Mancabelli L, Anzalone R, et al. Colonization of the human gut by bovine bacteria present in Parmesan cheese. Nat Commun. 2019;10(1):1286.

Publications

1. **Longhi G**, Argentini C, Fontana F, Tarracchini C, Mancabelli L, Lugli GA, Alessandri G, Ventura M, Turroni F, Milani C. **"Saponin treatment for eukaryotic DNA depletion alter the microbial DNA profiles by reducing the abundance of Gram-negative bacteria in metagenomics analyses".** Microbiome Research Reports. *Under revision.*

2. Alessandri G, Fontana F, Tarracchini C, Rizzo SM, Bianchi MG, Taurino G, Chiu M, Lugli GA, 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 Jan. (IF: 5.59).

3. Fontana F, **Longhi G**, Tarracchini C, Mancabelli L, Lugli GA, Alessandri G, Turroni F, Milani C, Ventura M. "**The human gut microbiome of athletes: metagenomic and metabolic insights".** Microbiome. 2023 Jan. (IF: 16.837).

4. Fontana F, **Longhi G**, Alessandri G, Lugli G.A, Mancabelli L, Tarracchini C, Viappiani A, Anzalone R, Ventura M, Turroni F, Milani C. "**Multifactorial microvariability of the Italian raw milk cheese microbiota and implication for urrent regulatory scheme**". mSystems. 2022 Dec. doi: 10.1128/msystems.01068-22 (IF: 7.324).

5. 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 infanti gut bacterial biodiversity by meta-analysis and formulation of an optimized cultivation medium".** NPJ Biofilms Microbiomes. 2022 Oct. doi: 10.1038/s41522-022-00349-1. (IF: 7.55)

6. Fontanta F, Alessandri G, Tarracchini C, Bianchi MG, Rizzo SM, Mancabelli L, Lugli GA, Argentini C, Vergna LM, 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 Sep. doi: 10.1111/1462-2920.16205. (IF: 5.491)

7. 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. doi: 10.1128/aem.00522-22. (IF: 4.792).

8. **Longhi G**, GA Lugli, 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 Mar. doi: 10.1111/1462-2920.15988. (IF: 5.491).

9. **Longhi G**, Lugli GA, 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. doi: 10.3389/fmicb.2021.790881. (IF: 5.59).

10. Tarracchini C, Milani C, **Longhi G**, Fontana F, Mancabelli L, Pintus R, GA Lugli, Alessandri G, Anzalone R, Viappiani A, Turroni F, Mussap M, Dessì A, Marincola FC, Noto A, De Magistris A, Vincent M, Bernasconi S, Picaud JC, Fanos V, Ventura M. **"Unraveling the microbiome of necrotizing enterocolitis: insights in novel microbial and metabolomic biomarkers".** Microbiol Spectr. 2021 Oct. doi: 10.1128/Spectrum.01176-21. (IF: 7.171).

11. Fontana F, Mancabelli L, GA Lugli, 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. doi: 10.1111/1758-2229.12960. (IF: 3.51).

12. Tarracchini C, Milani C, GA Lugli, 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. doi: 10.1099/mgen.0.000609. (IF: 4.868).

13. Milani C, Lugli GA, Fontana F, Mancabelli L, Alessandri G, **Longhi G**, Anzalone R, Viappiani A, Turroni F, van Sinderen D, Ventura **"M. METAnnotatorX2: a comprehensive tool for deep and shallow metagenomic data set analyses".** mSystems. 2021 Jun. doi: 10.1128/mSystems.00583-21. (IF: 6.53).

14. Mancabelli L, Mancino W, GA Lugli, Argentini C, **Longhi G**, Milani C, Viappiani A, Anzalone R, Bernasconi S, van Sinderen D, Ventura M, Turroni F. **"Amoxicillin-clavulanic acid resistance in the genus *Bifidobacterium*".** Appl Environ Microbiol. 2021 Mar. doi: 10.1128/AEM.03137-20. (IF: 4.792)

15. Mancabelli L, Mancino W, Lugli GA, Milani C, Viappiani A, Anzalone R, **Longhi G**, van Sinderen D, Ventura M, Turroni F. **"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. 2021 Feb. doi: 10.1128/AEM.02899-20. (IF: 4.792).

16. Neuzil-Bunesova V, Lugli GA, Modrackova N, Vlkova E, Bolechova P, Burtscher J, **Longhi G**, Mancabelli L, Killer J, Domig K, Ventura M**. "Five novel bifidobacterial species isolated from faeces of primates in two Czech zoos: *Bifidobacterium erythrocebi* sp. nov., *Bifidobacterium moraviense* sp. nov., *Bifidobacterium oedipodis* sp. nov., *Bifidobacterium olomucense* sp. nov. and *Bifidobacterium panos* sp. nov".** Int J Syst Evol Microbiol. 2021 Jan. doi: 10.1099/ijsem.0.004573. (IF: 2.4).

17. Duranti S, **Longhi G**, Ventura M, van Sinderen M, Turroni F. **"Exploring the ecology of Bifidobacteria and their genetic adaptation to the mammalian gut".** Microorganisms. 2020 Dec. doi: 10.3390/microorganisms9010008. (IF: 4.78).

18. Fontana F, Alessandri G, Lugli GA, Mancabelli L, **Longhi G**, Anzalone R, Viappiani A, Ventura M, Turroni F, Milani C. **"Probiogenomics analysis of 97 *Lactobacillus crispatus* strains as a tool for the identification of promising Next-Generation Probiotics".** Microorganisms. 2020 Dec. doi: 10.3390/microorganisms9010073. (IF: 4.78).

19. Milani C, Alessandri G, Mancabelli L, Mangifesta M, Lugli GA, Viappiani A, **Longhi G**, Anzalone R, Duranti S, Turroni F, Ossiprandi MC, van Sinderen D, Ventura M. **"Multi-omics approaches to decipher the impact of diet and host physiology on the mammalian gut microbiome".** Appl Environ Microbiol. 2020 Nov. doi: 10.1128/AEM.01864-20. (IF: 4.792)

20. **Longhi G**, van Sinderen D, Ventura M, Turroni F. **"Microbiota and cancer: the emerging beneficial role of bifidobacteria in cancer immunotherapy".** Front Microbiol. 2020 Sep. doi: 10.3389/fmicb.2020.575072. (IF: 5.59).

21. Duranti S, Ruiz L, Lugli GA, Tames H, Milani C, Mancabelli L, Mancino W, **Longhi G**, Carnevali L, Sgoifo A, Margolles A, Venura M, Ruas-Madiedo P, Turroni F. **"*Bifidobacterium adolescentis* as a key member of the human gut microbiota in the production of GABA".** Sci Rep. 2020 Aug. doi: 10.1038/s41598-020-70986-z. (IF: 4.996).

22. Milani C, Alessandri G, Mangifesta M, Mancabelli L, Lugli GA, Fontana F, **Longhi G**, Anzalone R, Viappiani A, Duranti S, Turroni F, Costi R, Annicchiarico A, Morini A, Sarli L, Ossiprandi MC, van Sinderen D, Ventura M. **"Untangling species-level composition of complex bacterial communities through a novel metagenomic approach".** mSystems. 2020 Jul. doi: 10.1128/mSystems.00404-20. (IF: 6.53).

23. Milani C, Fontana F, Alessandri G, Mancabelli L, Lugli GA, **Longhi G**, Anzalone R, Viappiani A, Duranti S, Turroni F, Ossiprandi MC, van Sinderen D, Ventura M. **"Ecology of Lactobacilli present in Italian cheeses produced from raw milk".** Appl Environ Microbiol. 2020 Jun. doi: 10.1128/AEM.00139-20. (IF: 4.792)

24. Lugli GA, 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 MI, Ossiprandi MC, van Sinderen D, Ventura M. **"Investigating bifidobacteria and human milk oligosaccharide composition of lactating mothers".** FEMS Microbiol Ecol. 2020 May. doi: 10.1093/femsec/fiaa049. (IF: 4.519).

25. Duranti S, Lugli GA, Viappiani A, Mancabelli L, Alessandri G, Anzalone R, **Longhi G**, Milani C, Ossiprandi MC, Turroni F, Ventura M. **"Characterization of the phylogenetic diversity of two novel species belonging to the genus *Bifidobacterium*: *Bifidobacterium cebidarum* sp. nov. and *Bifidobacterium leontopitheci* sp. nov".** Int J Syst Evol Microbiol. 2020 Apr. doi: 10.1099/ijsem.0.004032. (IF: 2.4).

26. Alessandri G, Milani C, Mancabelli L, **Longhi G**, Anzalone R, Lugli GA, Duranti S, Turroni F, Ossiprandi MC, van Sinderen D, Ventura M. **"Deciphering the bifidobacterial populations within the canine and feline gut microbiota".** Appl Environ Microbiol. 2020 Mar. doi: 10.1128/AEM.02875-19. (IF: 4.792)

27. Milani C, Alessandri G, Mancabelli L, Lugli GA, **Longhi G**, Anzalone R, Viappiani A, Duranti S, Turroni F, Ossiprandi MC, van Sinderen D, Ventura M. **"Bifidobacterial distribution across italian cheeses produced from raw milk".** Microorganisms. 2019 Nov. doi: 10.3390/microorganisms7120599. (IF: 4.78).

28. Mancino W, Duranti S, Mancabelli L, **Longhi G**, Anzalone R, Milani C, Lugli GA, Carnevali L, Statello R, Sgoifo A, van Sinderen D, Ventura M, Turroni F. **"Bifidobacterial transfer from mother to child as examined by an animal model".** Microorganims. 2019 Aug. doi: 10.3390/microorganisms7090293. (IF: 4.78).

29. Duranti S, Mancabelli L, Mancino W, Anzalone R, **Longhi G**, Statello R, Carnevali L, Sgoifo A, Bernasconi S, Turroni F, Ventura M. **"Exploring the effects of COLOSTRONONI on the mammalian gut microbiota composition".** PLoS One. 2019 May. doi: 10.1371/journal.pone.0217609. (IF: 3.240).

30. Lugli GA, Mancino W, Milani C, Duranti S, Mancabelli L, Napoli S, Mangifesta M, Viappiani A, Anzalone R, Longhi G, van Sinderen D, Ventura M, Turroni F. **"Dissecting the evolutionary development of the species *Bifidobacterium animalis* through comparative genomics analyses".** Appl Environ Microbiol. 2019 Mar. doi: 10.1128/AEM.02806-18. (IF: 4.792)

31. Lugli GA, Duranti S, Albert K, Mancabelli L, Napoli S, Viappiani A, Anzalone R, **Longhi G**, Milani C, Turroni F, Alessandri G, A Sela D, van Sinderen D, Ventura M. **"Unveiling genomic diversity among members of the species *Bifidobacterium pseudolongum*, a widely distributed gut commensal of the animal kingdom".** Appl Environ Microbiol. 2019 Apr. doi: 10.1128/AEM.03065-18. (IF: 4.792)

32. Milani C, Duranti S, Napoli S, Alessandri G, Mancabelli L, Anzalone R, **Longhi G**, Viappiani A, Mangifesta M, Lugli GA, Bernasconi S, Ossiprandi MC, van Sinderen D, Ventura M, Turroni F. **"Colonization of the human gut by bovine bacteria present in Parmesan cheese".** Nat Commun. 2019 Mar. doi: 10.1038/s41467-019-09303-w. (IF: 17.69).