UNIVERSITÀ DI PARMA

UNIVERSITÀ DEGLI STUDI DI PARMA

Dottorato di ricerca in Biotecnologie e Bioscienze

XXXVI Ciclo

Geography, diet and host's lifestyle shape the human gut microbiota

Coordinatore:
Chiar.mo Prof. Marco Ventura

Tutor:
Chiar.ma Prof.ssa Francesca Turroni

Dottorando:
Federico Fontana

Parma, 2020/2021 – 2022/2023

Table of Contents

Summary 5

Chapter 1 General Introduction 7

A. Gut Microbiota: a dynamic bacterial population. 8

B. Factors shaping the human gut microbiota 11

Host health status, drug treatment, and gut microbiota modulation 11

Dietary variability and the influence of fermented foods in shaping the gut microbiota composition 14

Age and host's lifestyle correlate with the gut microbiota composition 16

Environment as a source of microbial variability that shapes the human gut microbiota. 19

C. New tools for in-depth microbiota analysis. 21

Cutting-edge innovations in DNA processing and sequencing 23

Advancement in Culture-independent approaches for the profiling of complex microbial communities 25

Chapter 2 Outline of the Ph.D. thesis. 30

Chapter 3 Investigation of the Ecological Link between Recurrent Microbial Human Gut Communities and Physical Activity 33

Chapter 4 The human gut microbiome of athletes: metagenomic and metabolic insights 59

Chapter 5 Investigating the infant gut microbiota in developing countries: worldwide metagenomic meta-analysis involving infants living in sub-urban areas of Côte d'Ivoire. 91

Chapter 6 Multifactorial Microvariability of the Italian Raw Milk Cheese Microbiota and Implication for Current Regulatory Scheme 120

Chapter 7 Designation of optimal reference strains representing the infant gut bifidobacterial species through a comprehensive multi-omics approach. . 158

Chapter 8 General Conclusion..... 197

References 202

Publications in peer-reviewed journals achieved in the course of the Ph.D. 217

Summary

The microbiota, defined as the collection of bacteria inhabiting a specific ecological niche, is an expanding field of research. Microbiota itself is ubiquitous and significantly impacts a wide range of domains, including human beings as well as all the currently known environments. Recent research endeavors have highlighted the pivotal role played by bacterial communities in human lives, which is manifested through direct and indirect interactions. Nevertheless, despite sharing similar environmental conditions, the microbiota composition associated with a specific niche usually shows a high degree of variability, and this biodiversity is the source of various translational trial applications.

The gut microbiota, among the many types of human-related microbiota, has caught the interest of researchers due to its deep complexity and high level of interaction with the human host. Actually, the human gut microbiota is regarded as one of the key variables that may affect the outcome of oral drug treatment, as well as metabolism and even brain functionality with psychological implications. For these reasons, comprehensive investigations of the gut microbiota composition in infants and adults led in the past decades to the identification of specific recurring bacterial communities in healthy subjects while correlating some bacterial taxa with the onset of specific disorders and the presence of chronic diseases.

These correlations between the bacterial composition of the human gut and health status highlight the potential outcomes that could be achieved by exploiting the enhanced availability of gut microbiome data. Overall, studying the correlation between bacterial prevalence, abundance and correlation with the human host's health status now offers valuable insights into the role of the human microbiota

that will drive the progress of personalized medicine supporting human well-being.

The underlying basis of my Ph.D. lies in examining microbial compositions associated with the human gut and accompanied by specific metadata such as age, geographical location, and lifestyle, as well as developing novel bioinformatic tools for data analysis. Furthermore, given that diet is one of the most complex factors capable of influencing the composition of the human gut, a thorough investigation into the role of fermented foods in delivering bacterial strains to the human gut was conducted.

Chapter 1

General Introduction

A. Gut Microbiota: a dynamic bacterial population.

Microbiota is the collective microbial community that inhabits a specific environment [1]. Each environment, including soil, water, food, insect guts, and the different districts of the human body, harbors a unique microbiota composition [2-5]. This specific and unique microbiota arrangement is due to bacteria's ability to adapt and withstand different environmental pressures and thrive in different ecological niches. Indeed, an ecological niche with a highly variable microbiota is the human gastrointestinal tract (GIT) [6]. The composition of the so-called gut microbiota is constantly shaped by a variety of factors, both external (e.g., diet, lifestyle) and host-related (e.g., genetics, immune system, stress) [7]. As a result of the pressure exerted by these variables, the gut microbiota can shift into a homeostatic equilibrium condition characterized by high compositional stability and resilience, or in a dysbiotic state, characterized by high variability and compositional instability [8-10]. Since the earliest timing of life, the gut bacteria establish symbiotic interactions with the host, promoting the development of an effective immune system [11-13]. In contrast, the condition of an imbalanced microbiota composition, known as dysbiosis, has been frequently associated with negative impacts on human health [14,15]. In this context, a balanced and healthy diet, the absence of chronic diseases, low-stress levels, and a healthy lifestyle are all factors that support a balanced state of the gut microbiota, which is generally associated with improved human health status [16]. Conversely, diets comprising highly processed foods with excessive salt, fat, and animal protein, along with

chronic diseases and the intake of antibiotics, may disrupt the balance of gut microbiota [15].

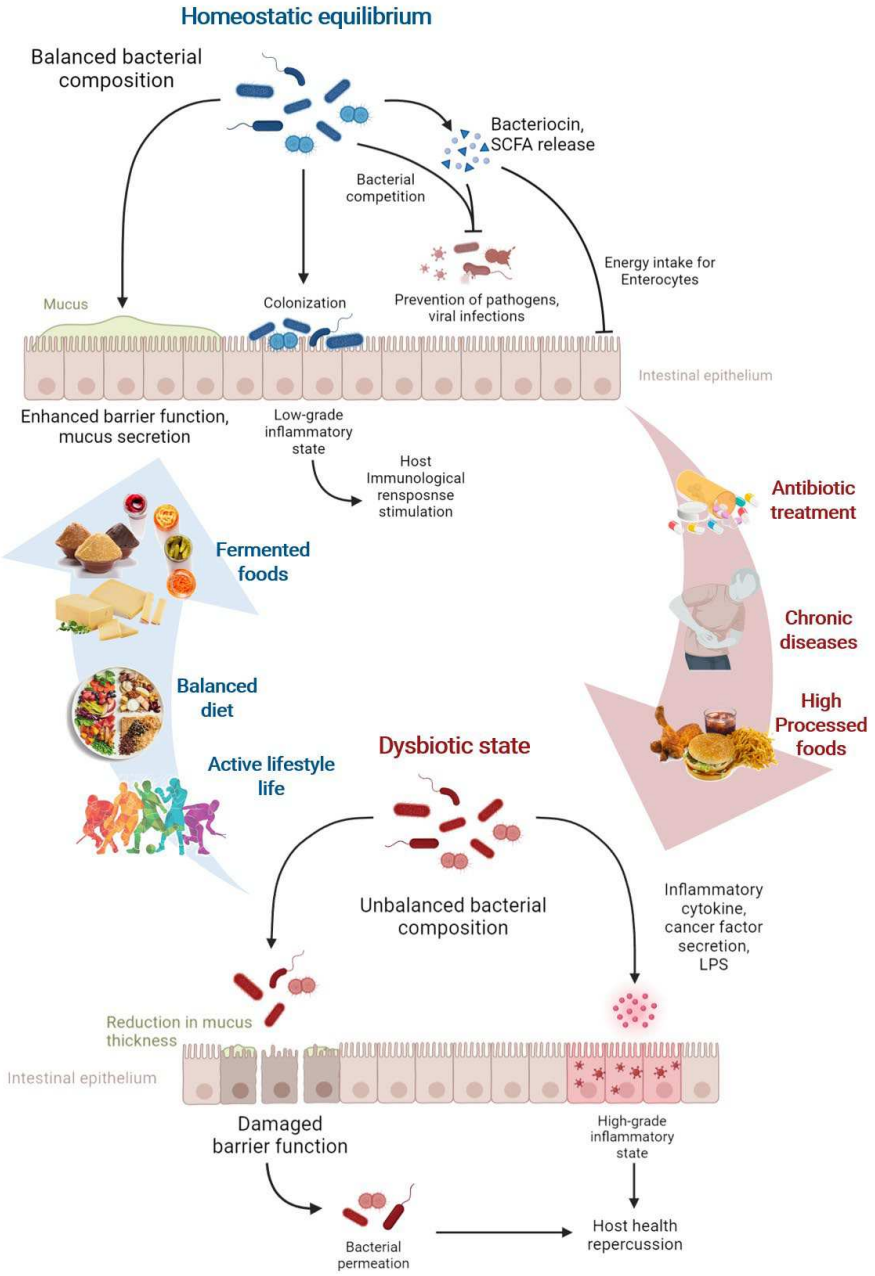


Figure 1. Striking a delicate equilibrium in the human gut: dysbiosis versus homeostasis. This figure illustrates the dynamic relationship between the gut microbiota and human intestinal

cells, depicting both the stable state of bacterial equilibrium (homeostatic balance) and the imbalanced state (dysbiotic state). The figure also illustrates several factors that promote the directional transition between the two states of the human gut microbiota.

Remarkably, different microbiota taxonomic compositions have been observed to possess distinct metabolic and functional capacities, allowing for unique interactions with the human host [17]. For instance, a gut microbiota dominated by microorganisms producing Short-Chain Fatty Acids (SCFAs), such as *Faecalibacterium*, *Ruminococcus*, and *Eubacterium* genera, may improve the energy availability for the human intestinal cell, the enterocytes, and promote T cells expansion [18,19]. The integration of compounds with high biological value, such as vitamins (menaquinone, folate, biotin, riboflavin, etc.), is also associated with the biological activity of gut microbiota other than diet alone [19,20]. Intriguingly, certain bacteria can also metabolize multiple drugs that reach the gut through oral consumption, reducing their impact and the effectiveness of the pharmacological treatment [21].

Therefore, understanding the metabolic and functional properties of the intestinal microbiota, in addition to its taxonomic composition, is of critical importance for the improvement of human health-related aspects.

B. Factors shaping the human gut microbiota

The human-derived microbiota and the human host coexist, interacting each other in both positive and negative ways [6,22,23]. Thus, understanding the factors that influence the gut microbiota can help us to prevent diseases caused by the dysbiosis of the gut microbiota, as well as improve the efficacy of drug-mediated treatment, and increase the production of beneficial compounds in the gut lumen [10]. The main factors that can influence the stability of the gut microbiota by shaping its microbial composition and impacting its proper balance will be discussed in this section.

Host health status, drug treatment, and gut microbiota modulation

Disorders in the gut microbiota can lead to higher levels of inflammation in the gut, which may induce a reversible inflammatory condition like irritable bowel syndrome (IBS). IBS is a common condition that impacts a large number of people and is influenced by several factors, such as the disruption of gut bacteria homeostasis caused by long-term stress or an inadequate diet [24,25]. However, the dysbiotic-related condition of the gut microbiota can also be affected by chronic diseases [26]. This is the case of inflammatory bowel diseases (IBD) such as

Crohn's disease and ulcerative colitis, which are complex and multifactorial pathological conditions primarily related to the host's genetics involving high levels of chronic digestive tract inflammation [27,28]. Nevertheless, the gravity of IBD symptoms can depend highly on the resident's gut microbiota composition, which can increase the severity of the chronic inflammatory state by overstimulating the natural immune response [26,29].

Furthermore, different bacterial strains are associated with distinct drug-related metabolisms, which can alter orally administered drugs and reduce the prodrugs' intestinal absorption [30]. An example is *Fusobacterium nucleatum,* which is often found on the surfaces of colorectal tumors due to the optimal growth conditions represented by the physiologically altered environment [31]. Clinical research on colorectal cancer-affected patients showed that some strains of *Fusobacterium nucleatum* can degrade the most used chemotherapeutic drugs, thus reducing the pharmacological response to chemotherapy [32]. Thus, obtaining a comprehensive and optimal understanding of the microbial composition associated with the health or pathological status of the human host is crucial for preventing the onset of diseases and enhancing the overall effect of the medical treatment.

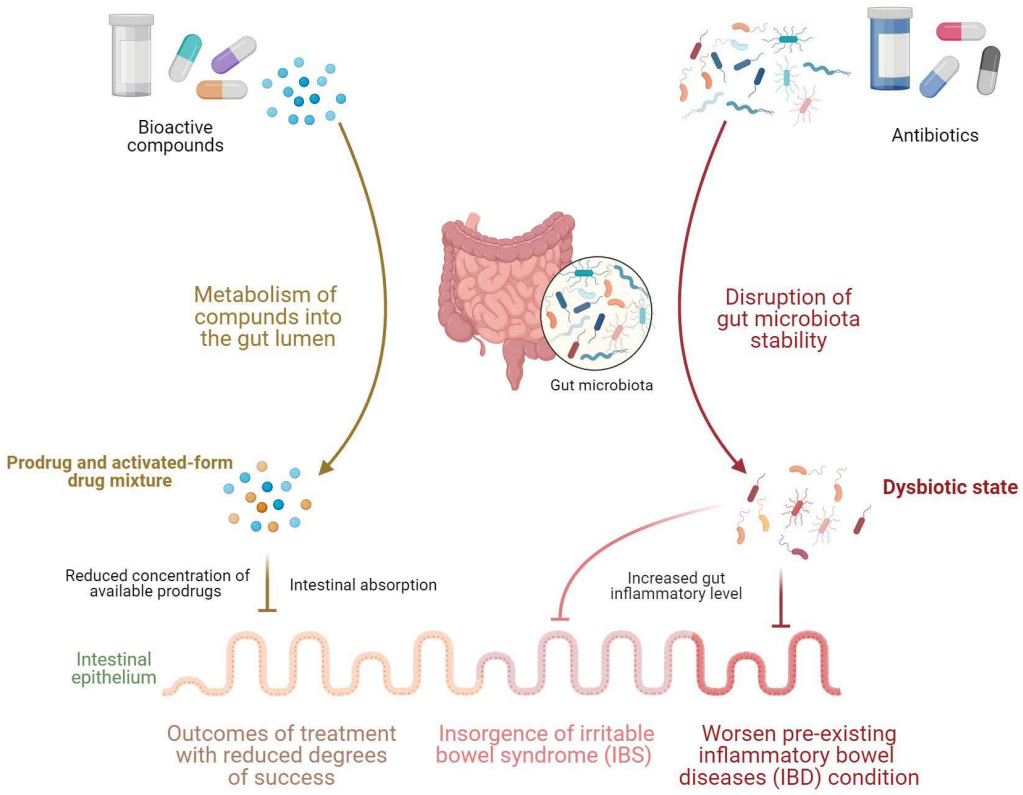


Figure 2. Investigating the impact of drugs on the gut microbiota and its implications for overall health. This figure illustrates the detrimental effects of antibiotics on the gut microbiota, leading to an imbalance in its composition. In addition, the figure highlights the impact of gut microbiota's metabolic activity on the efficacy of pharmaceutical treatments. It also shows how the dysbiotic's microbiota activity can contribute to increased inflammation of the intestinal barrier, potentially exacerbating pre-existing conditions like IBD.

Furthermore, oral administration of a broad-spectrum of antibiotics to treat infections in different human body districts can shift the gut microbiota homeostasis into a dysbiotic condition [33]. In most cases, short-term antibiotic treatments result in a transient state of dysbiosis that spontaneously resolves due to the resilience of the intestinal microbiota [34,35]. However, prolonged antibiotic

treatments can irreversibly alter the climax of the gut microbiota, promoting the development of more resistant strains and permanently shifting the bacterial gut balance into an altered condition [33]. Additionally, the use of non-antibiotic drugs can also alter the gut microbiota stability [36].

Thus, to prevent excessive disruptions of the gut microbiota caused by long-term antibiotic use, incorporating probiotic supplements and maintaining a balanced and well-rounded diet can be beneficial.

Dietary variability and the influence of fermented foods in shaping the gut microbiota composition.

Diet is one of the main drivers in establishing a stable and balanced intestinal microbiota [37,38].

Foods can be classified as low-processed or highly-processed. Consumption of low-processed foods with low fat and salt content, as well as a well-balanced combination of protein and fiber derived from vegetables, has been linked to increased Prevotellacee and *Bacteroidetes* taxa [39]. On the other hand, consuming an elevated quantity of highly processed foods is associated with increased animal protein and lipids intake, which can result in higher levels of *Fusobacterium* and other mucin-degrading bacteria [40-42]. Thus, radical changes in the diet can modulate the composition of the gut microbiota by altering the organic substrate that reaches the human gut.

While the gut microbiota can be influenced by changes in diet at any age, the most significant shifts in the microbial composition that correlate with age and diet occur during the weaning period [43]. The transition from a milk-only (liquid) to a multi-source (solid) diet induces a significant shift in the predominant bacteria found in the large intestine of infants [44]. As a result, *Bifidobacterium* (and other Actinobacteria) start to decrease over time, while Firmicutes and Bacteroidetes start to increase [45]. In detail, the gut-related bifidobacterial population changed substantially with *Bifidobacterium longum subsp. longum* and *Bifidobacterium adolescentis* that emerged as the new dominant bifidobacterial species after weaning [46]. Instead, pre-weaning supporting bacterial genera such as *Bacteroides* and *Veillonella* start to arise as newly prevalent gut microbiota members [45].

Diet can also directly modulate the gut microbiota composition by consuming fermented foods and probiotic supplements, with significant health benefits for the human consumer [47]. Fermented foods like yogurt, cheese, kefir, and other products derived from fermented milk are widely consumed around the globe and are a natural source of beneficial bacterial cells [48]. In detail, the food-related microbiota, commonly dominated by LAB, modifies the organoleptic properties of fermented products by releasing different compounds [49]. Fermentation is a food-production technique that has been used for centuries to enhance and preserve various food products, preventing spoilage and ensuring consistent sensory properties [50,51]. Lactic acid bacteria, in particular, contribute to the acidification of the fermented product, resulting in changes in taste, texture, and composition with high technical and industrial repercussions [52]. This acidification also plays an important role in reducing the overall bacterial load,

effectively preventing the growth of bacterial species that may be contaminated in the environment [53,54]. This microbial-derived refinement of fermented products turns them into “functional foods” that can provide biologically high-value molecules into the diet, such as vitamins and other byproducts of bacterial metabolism as well as living bacterial cells [55-57].

For these reasons, the pivotal role of food microbiota in modulating human gut microbiota is unsurprising, given its regular consumption through the typical diet. Consequently, delving into the effects of diets on the microbial communities within the human gut becomes imperative, offering insights that could facilitate the identification of specific dietary approaches to effectively shape and promote the desired gut microbiota composition.

Age and host’s lifestyle correlate with the gut microbiota composition.

Due to the influence of many factors that compose and shape the lifestyle of individuals, the gut microbiota constantly evolves and changes in relation to the human host's age [58]. Competitive sports, for example, demand a very different lifestyle than sedentary people, including different diets and levels of physical and mental stress. These lifestyle changes have been shown to significantly impact the gut microbiota composition, particularly according to recent studies that have revealed extremely high biodiversity among healthy sedentary and athletic individuals [59]. From a metabolic point of view, professional athletes also showed the dominant presence of SCFAs bacterial producer species, whereas

sedentary subjects had an increased presence of *Bacteroides* [60]. As exemplified by the difference between athletes and sedentary individuals, lifestyle represents a major set of environmental factors shaping the gut microbiota composition.

Another key factor is represented by aging. Indeed, distinct age groups can be identified by their recurrent microbial profiles, which are classified as Community State Typers (CSTs), dominated by age-related bacterial taxa [61]. These CSTs are associated with different age groups, including infancy, adolescence, adult, and elderly, which showed recurring and commonly shared microbial taxa [61]. In infancy, the gut microbiota is commonly associated with a healthy state, i.e., gut microbiota into a homeostatic equilibrium, and exhibits the dominance of species belonging to the *Bifidobacterium* genus [62]. Bifidobacterial species like *Bifidobacterium bifidum*, and *Bifidobacterium breve* are among the most relevant bacterial species in healthy infants under the first year of life [63]. These bacteria are genetically well-suited for using the HMOs (human milk oligosaccharides) and other substrates present in breast milk, allowing them to colonize the gastrointestinal tract of infants more efficiently than other bacterial competitors [64]. Indeed, a recent metagenomics analysis performed on a large pool of infant samples under one year of life showed the presence of well-defined Infant Community State Types (ICSTs) representing recurring microbial profiles in the gut microbial population [65]. In contrast, unhealthy children with food or respiratory allergies display dysbiotic gut microbiota, with *Ruminococcus gnavus* being one of the main allergy markers [66].

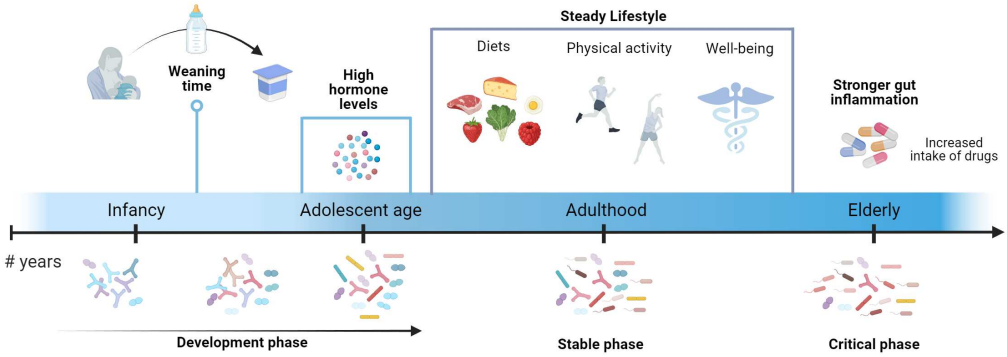


Figure 3. Factors influencing the composition of the human gut microbiota as individuals age. This figure illustrates the various stages of gut microbiota stability throughout an individual's lifespan and the key factors contributing to its variability and modulation during critical life stages.

Instead, during childhood and adolescence, multiple factors like family-related diet, individual physical activity, and lifestyle choices impact the gut microbiota composition in humans, resulting in a wide range of gut microbial variations [67]. Then, as a long-term effect of lifestyle, the gut microbial composition tends to reach a stable condition in healthy adults [68]. In detail, gut microbiota composition in healthy adults becomes relatively stable, with bacterial genera such as *Bacteroides*, *Veillonella*, *Faecalibacterium*, *Prevotella,* and *Ruminococcus* among the most representative ones [68].

Finally, due to physiological changes in the host's health, immune dysregulation, and pathologies associated with aging, the gut microbiota changes in seniors [69]. The balance of the gut microbiota in senior subjects is critical since it is frequently prone to increase pro-inflammatory bacterial species, which can cause chronic inflammation, resulting in a worsening of the host's general well-being [70].

All these data highlight how getting insight into the CSTs associated with each age group and health status is essential for preventing the onset of chronic inflammation and promoting well-being throughout life.

Environment as a source of microbial variability that shapes the human gut microbiota.

From a long-term perspective, the composition of the human gut can be affected by a range of environmental factors, such as humidity, hygienic conditions, and exposure to sources of contamination both in outdoor and indoor spaces [71].

This subject is crucial since current research shows that the environment overcomes host genetics in developing human gut microbiota [72].

Pets like cats and dogs are one of the primary sources of environmental contamination as they live in close contact with the family units, increasing the likelihood of interaction with bacterial contamination [73]. Furthermore, considerable scientific literature has highlighted how bacterial species residing within the human oral cavity and gut are shared between different family members, although the underlying mechanisms are still unclear [74,75]. Overall, the environment is a complex combination of factors that influence the gut microbiota, such as the transmission of living bacteria from the environment or the host's physiological response to high stress levels associated with challenging conditions [76].

Additionally, the environment plays a crucial role in shaping the fermented food microbiota, which ultimately may impact the gut microbial composition of consumers. Furthermore, fermented products' final flavor and texture, such as cheese, are primarily determined by the microbial community that develops during fermentation [77-80]. Among the main bacterial taxa in the cheese microbiota, Lactic acid bacteria (LAB) competitively dominate over other bacterial taxa [81,82]. However, factory-related microbial sources of contamination often influence the final food microbiota composition [83]. Indeed, utilizing fermented material from past fermentation cycles is a traditional artisanal method that can yield products with distinct and complex organoleptic features. This method employs Natural Whey Cultures, i.e., unique microbial compositions derived from the fermented material retained from previous cheese production rounds [84,85]. These NWCs microbial starters are unique to each food processing site, naturally shaped by the unique environmental pressure and production process [85]. Consequently, the careful selection and management of LABs used in food production, such as the use of Natural Whey Culture (NWC) with unique microbiota composition or the use of ad-hoc microbial starter, is essential for developing high-quality products [86,87].

Thus, it is essential to note that environmental factors can influence the food microbiota, which constitutes one of the primary drivers of microbial diversity in fermented products, resulting in distinct sensory characteristics and flavors while also vetting such microbes to the gastrointestinal tract of consumers.

C. New tools for in-depth microbiota analysis.

Throughout the last few years, several efforts have been made in order to disclose the "dark matter" of microbiota complexity in various environments, including the development of novel techniques, resources, and approaches. The concept of "dark matter" refers to the challenging microbial complexity undetectable from traditional cultivation methods and thus remained a mystery until the advent of culture-independent techniques [88]. The multidisciplinary nature of these efforts has become evident in both laboratory work and bioinformatics routines, with the development of more efficient protocols for sample processing and customized informatics pipelines for data analysis [89]. In addition, the development of II° and III° generation of DNA sequencing techniques, called next-generation DNA sequencing (NGS) technologies, has led to a significant increase in the volume of sequence data that could be accessible at optimal quality/quantity/price ratios [90]. Given the increased availability of microbiological data from DNA sequencing, an enormous amount of data is now available to the scientific and research communities in online databases (e.g., NCBI, EMBL) [91]. As a result, culture-independent approaches such as bacterial genomics, comparative genomics, and metagenomics have flourished, providing a large amount of data capable of driving culture-dependent techniques.

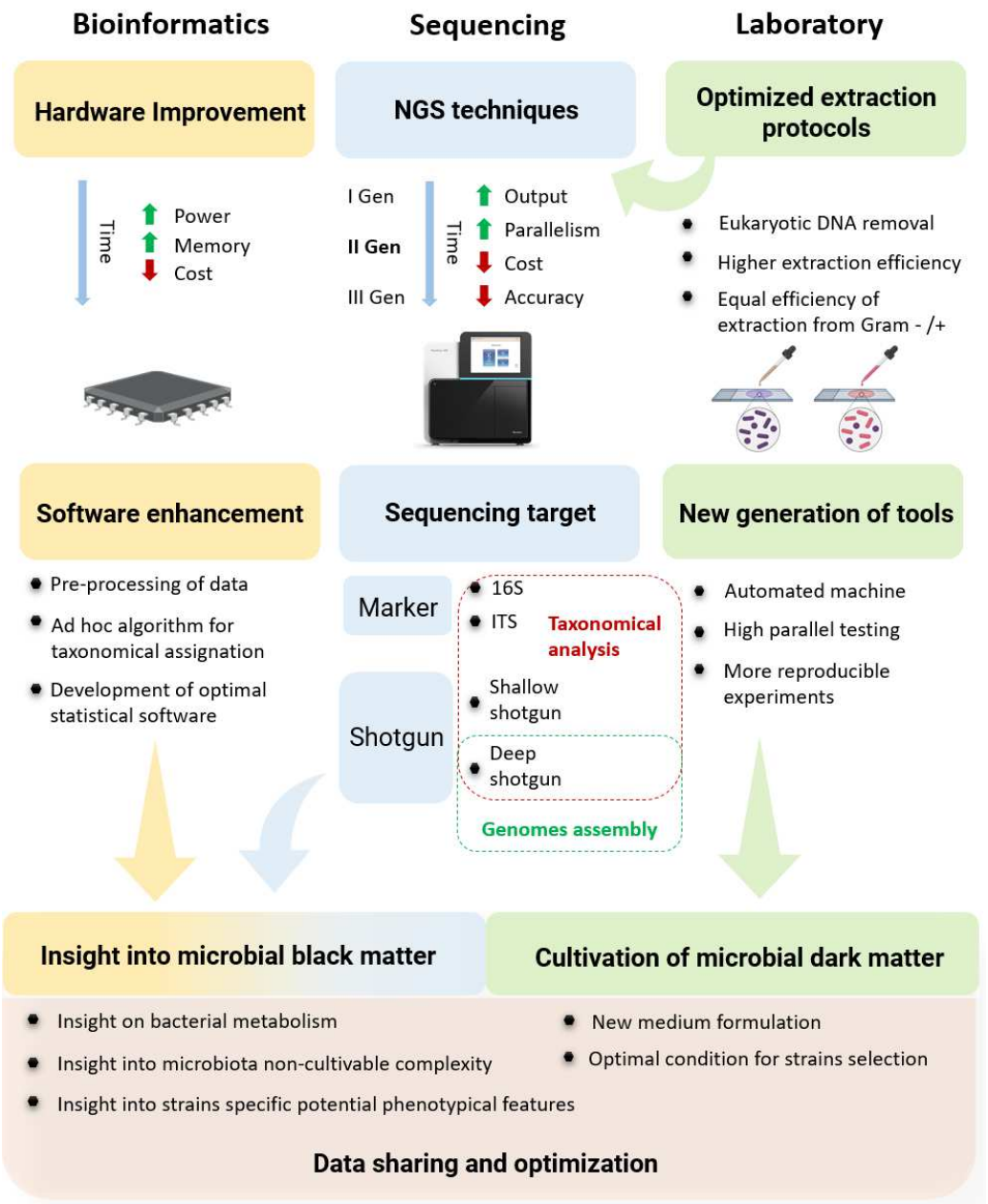


Figure 4. Major advancements in the three domains of contemporary microbiological research. The figure illustrates the latest bioinformatics, sequencing, and laboratory technology advancements. Additionally, the image highlights the interconnectedness of the three areas, enabling the attainment of previously unreachable goals.

Cutting-edge innovations in DNA processing and sequencing

Genomics and metagenomics have witnessed remarkable advancements in recent years, driven by the development of more efficient, cost-effective, and high-performing DNA extraction, manipulation, purification, and sequencing tools [92]. The extraction of high-quality bacterial DNA is the first critical step in processing the prokaryotic DNA. Complex biological sources like biological samples with high eukaryotic DNA contamination have long been challenging in prokaryotic DNA extraction [93]. Due to the isolation of DNA from complex biological matrix, traditional extraction methods often suffer from contamination issues, as well as by selective DNA extraction from different bacterial taxa, and lower extraction kit efficacy [94]. However, refinements in the DNA extraction and DNA purification protocols have dramatically improved over time, allowing high-quality procaryotic DNA extraction even from samples with low bacterial load [95].

Early DNA sequencing techniques based on the Sanger approach [96], also indicated as I° generation sequencing, allowed us to obtain a low output of sequencing data with a high accuracy level [97]. Nevertheless, although Sanger proved to be highly effective in achieving accurate gene sequencing, a significantly higher amount of DNA is required to encompass the full range of microbiota variability in complex metagenomic samples.

Thanks to the advent of NGS technologies, genomics and metagenomics rapidly became the gold standard for assessing the genetic repertoire of microbial strains

and whole microbial populations, respectively [98,99]. In particular, the transition from the I° generation to the II° generation of DNA sequencing technologies has prompted a significant increase in sequencing data output due to extensive parallelization of analysis, increased efficiency, and a high level of automation [100]. Moreover, III° generation of DNA sequencing techniques possess other notable advantages, such as the absence of pre-amplification of target DNA, minimal DNA amount requested, and increased read length [101]. However, III° of DNA sequencing generation techniques also have a considerable error rate due to the more advanced technology and detectors required [102]. On the other hand, the II° generation of DNA sequencing, despite some drawbacks (requiring.g., DNA pre-amplification and producing short reads), has emerged as the most efficient DNA sequencing technology regarding data output and accuracy ratio [103]. Then, with the widespread application of advanced II° generation DNA sequencing methods, often in conjunction with cutting-edge III° generation techniques [104], it has become feasible to reconstruct high-quality bacterial genomes and precisely define complex bacterial populations, allowing the identification of previously undetectable bacterial species.

Advancement in Culture-independent approaches for the profiling of complex microbial communities

With the implementation of new protocols and novel DNA sequencing technologies, a vast amount of procaryotic genetic-related data is now accessible for analysis [105]. As a result, more efficient bioinformatics tools, refined databases, and greater computing resources have become critical in managing this vast data flow [106,107]. Server and computer performance advancements have skyrocketed over the years, allowing a broader range of researchers to use high-end terminals with less economic investment [108]. Additionally, local networks of computational servers and data networks shared with high-performance centers have increased computing power accessibility by providing university research centers with public computing nodes [109]. However, *in-silico* analyses require an optimal database in order to obtain high-level taxonomical profiling and genome reconstruction [110,111]. Public databases, such as SILVA for 16S, were among the first to be established, growing in size and becoming more accurate by constantly removing wrongly classified 16S DNA sequences [112].

The availability of genetic data, expanding genomic databases, and powerful computational resources, sequencing, and bioinformatic methodologies have allowed us to enter into a new era of microbiome exploration.

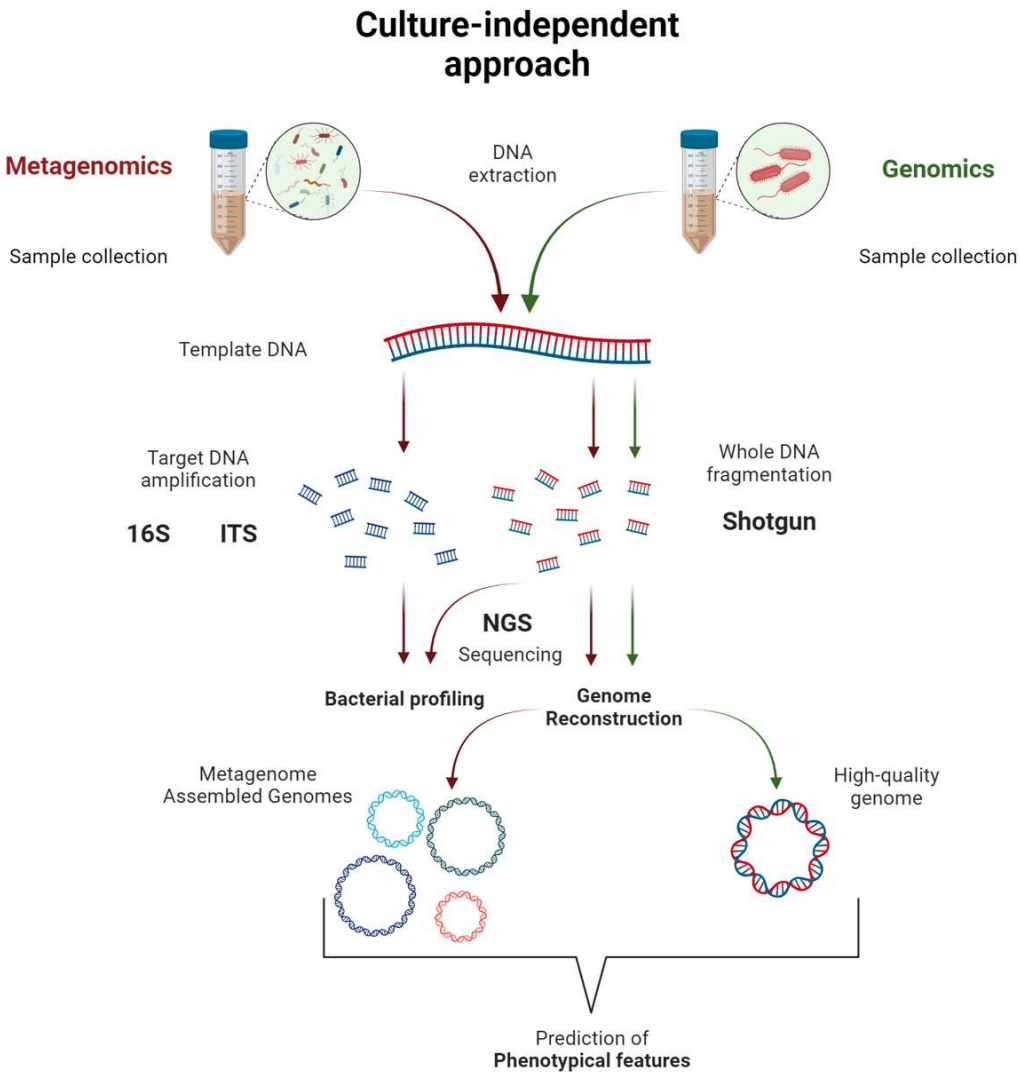


Figure 5. Culture-independent approach and its practical applications. The illustration depicts the crucial role of culture-independent techniques in metagenomic and genomic applications, enabling the reconstruction of bacterial genomes and the profiling of complex microbial populations.

Using II° generation DNA sequencing, prokaryotic profiling initially relied on amplifying marker sequences with ad-hoc primers, which is a relatively low-cost approach requiring low bioinformatics computational power [112]. The first marker

DNA sequencing approach focused on amplifying, sequencing, and analyzing the 16S ribosomal gene [113]. This genomic region is highly conserved and characterized by a slow mutation rate, making it the golden standard molecular clock and thus allowing the taxonomic identification of bacterial genera with high accuracy [114,115]. Alternative genetic markers with high genetic variability, such as the Internal Transcribed Spacer, or ITS, were later developed to reconstruct microbial composition up to the species level [116]. As an applicative example, ITS marker detection was extensively used to get insights into the composition of bifidobacterial communities [116,117]. However, these marker-based techniques also display significant limitations, such as the required primer-based pre-amplification step that may selectively affect the relative abundance of different bacterial genera [118,119].

During the last decade, many laboratories started to prefer shotgun techniques over marker-based techniques due to the advancements in computing power and upon the decrease of DNA sequencing costs [120]. By fragmenting and sequencing all the extracted microbial DNA, shotgun techniques allow the detection of bacterial taxa with species accuracy without the biases introduced by PCR amplification [121]. The shotgun technique requires significant computational power but allows us to overcome the limitations of the markers' gene-based techniques and provides insights into microbial populations' genetic repertoire [122]. This deep-sequencing analysis technique can also be applied to reconstruct metagenome-assembled genomes (MAGs), which are bacterial genomes reconstructed using specific bioinformatic pipelines from complex metagenomic samples rather than pure isolates [123]. Moreover, the MAGs reconstruction is commonly improved by coupling shotgun-derived short-reads with long reads

from III° generation techniques like Nanopore and PacBio to assemble high-quality genomes [124]. Instead, for high-quality taxonomy classification without the need to reconstruct MAGs, a less in-depth shotgun analysis, known as shallow shotgun, is often employed as an improved alternative to the 16S microbial profiling approach, requiring a lower economic investment [125].

Overall, the ability to highlight the bacterial complexity that was previously impossible to elucidate only with traditional bacterial cultivation methods started a new era in human microbiota investigation.

Chapter 2

Outline of the Ph.D. thesis.

The objective of this Ph.D. thesis is to provide novel insights into the main factors that affect the composition of the human gut microbiota. Thus, this Ph.D. thesis encompasses in-depth analysis of massive metagenomic datasets to thoroughly investigate the impact of athlete-related lifestyle and geographical localization on the composition and functionality of the human gut microbiota. Furthermore, this Ph.D. thesis investigated the impact of diet, specifically how fermented foods can act as a natural carrier of bacteria, ultimately influencing the human gut microbiota composition. In this regard, the microbiota composition of a large number of Italian raw milk cheeses has been investigated with great detail to gain insights into the microbial communities that they carry. Finally, we developed a bioinformatic-based approach to obtain valuable representative bacterial strains for each bacterial taxon encompassing the human gut microbiota based on an ecological and genetic point of view.

Overall, this thesis explores the main factors that modulate the human gut microbiota and the food microbiota, with implications for host health and food product quality.

Chapter 3 presents a taxonomic-oriented metagenomic analysis of 207 gut microbiota samples from healthy athletes and sedentary controls. The study aimed to find a correlation between physical activity, athletic lifestyles, and the modulation of gut microbiota, potentially impacting competitive performance.

Chapter 4 provides insight into the impact of lifestyle on the human gut microbiota composition. In detail, this study analyzed 418 shotgun metagenomics datasets from fecal samples of healthy athletes and sedentary adults to investigate the effects of agonistic sports and related lifestyles. These findings highlighted the existence of correlations between competitive athletes

and gut microbiota modulation, potentially leading to beneficial effects on human physical performance and health conditions.

Chapter 5 explores the impact of geographical location on the infant gut microbiota and its composition. The study exploited a dataset of 1098 shallow shotgun infant gut microbiota samples, including 11 fecal samples from Côte d'Ivoire de novo sequenced in our lab. The samples were divided into a simplified geographical denomination in rural and urban countries. The meta-analysis was performed to identify Infant Species Community State Types and specific microbial species covariances related to the geographical origins.

Chapter 6 reports a study regarding the taxonomical characterization of 128 Italian raw milk cheese samples and how the resident microbiota is linked to each cheesemaking site and site-specific Natural Whey Cultures (NWCs) rather than cheese type or geographical origin. These data shed light on the microbial population commonly ingested through cheese consumption.

Chapter 7 introduces a novel bioinformatic approach for identifying model bacterial bacterial taxa of the human gut microbiota. Specifically, we applied this *in silico* pipeline to *Bifidobacterium* genus, leading to the identification of model strains for the *B. bifidum*, *B. breve*, *B. longum* and *B. adolescentis* species. Remarkably. Such bifidobacterial taxa are key members of the infant gut microbiota and thus the identification of a model strain for each of these species represents an important achievement for the evaluation of the microbe-microbe and microbe-host cross-talks.

Chapter 3

Investigation of the Ecological Link between Recurrent Microbial Human Gut Communities and Physical Activity

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

The results of this chapter were published in Microbiology Spectrum, 2022 Apr 4; https://doi.org/10.1128/spectrum.02733-21.

[*] These authors contributed equally.

Abstract

Emerging evidence has shown an association between the composition of intestinal microbial communities and host physical activity, suggesting that modifications of the gut microbiota composition may support training, performance, and post-exercise recovery of the host. Nevertheless, investigation of differences in the gut microbiota between athletes and individuals with reduced physical activity is still lacking. In this study, we performed a meta-analysis of 207 publicly available shotgun metagenomics sequencing data of fecal samples from athletes and healthy non-athletes. Accordingly, analysis of species-level fecal microbial profiles revealed three recurring compositional patterns, named HPC1 to 3, that characterize the host based on their commitment to physical activity. Interestingly, the gut microbiome of athletes showed a higher abundance of anti-inflammatory, health-promoting bacteria than that of non-athletic individuals. Moreover, the bacterial species profiled in the gut of professional athletes are short-fatty acid producers, which potentially improve energy production, and therefore sports performances. Intriguingly, microbial interaction network analyses suggested that exercise-induced microbiota adaptation involves the whole microbial community structure, resulting in a complex microbe-microbe interplay driven by positive relationships among the predicted butyrate-producing community members.

Importance

Through metagenomic analyses, this work revealed that athletes have a gut-associated microbial community enriched in butyrate-producing species compared with non-athletes. This evidence can support the existence of a two-way association between the host's lifestyle and the gut microbiota composition, with potential intriguing athletic performance outcomes.

Introduction

The human gut harbors a complex community of microorganisms, commonly referred to as human gut microbiota, which is well-known to play a role in nutrient uptake, vitamin synthesis, energy harvest, inflammatory modulation, and host immune response (1,–3). In turn, numerous host-dependent factors, such as genetics, age, antibiotic use, and diet, can affect the gut microbiota resulting in a highly dynamic and individual gut ecosystem (4). Recently, it has been argued that physical activity can influence gut microbiota composition, depending on the type, intensity, and exercise duration. The gut microbiota, in return, may affect the athlete's health and performance (5). Indeed, if moderate exercises (50% to about 70% of the maximum heart rate) (6) have been reported to increase the overall gut microbiota's (bio)diversity (7), prolonged endurance exercises (70% to about 85% of the maximum heart rate) (6) have been linked with an increased abundance of gut bacterial species producing short-chain fatty acids (SCFAs) (8). In particular, members of the *Veillonella* genus, along with the metabolic pathways that this taxon utilizes for lactate conversion to propionate, have been detected with elevated abundances in athletes (9), thereby contributing to host metabolic efficiency by increasing energy availability, and thus ultimately

influencing athlete performance (10). Moreover, a recent study involving professional and competitive unprofessional cyclists showed that a high training load of the cyclists corresponds to a high abundance of gut-associated *Prevotella* genus members (11). Notably, the presence of this genus has been correlated with increased metabolism of branch chain amino acids, i.e., leucine, valine, and isoleucine (11), which stimulates muscle protein synthesis and accelerates recovery (12). Furthermore, athletes generally consume higher energy diets than sedentary individuals, maintaining a high consumption of carbohydrates and proteins and a low-fat intake, with implications in gut microbiota composition (13).

In this context, our study aimed to explore the microbial communities inhabiting the gut of athletes and non-athletic individuals to highlight compositional and structural differences at the species level. For this purpose, we performed a meta-analysis employing 207 shotgun metagenomics data sets retrieved from public repositories.

Results and Discussion

Meta-analysis of athletic and non-athletic individuals: data set selection and bioinformatics.

Public repositories were screened for all available shotgun metagenomic data sets of the gut microbiomes of the athletes and non-athletic individuals. Specifically, we selected fecal metagenomics data from multiple sources to avoid the limitations of a single-center study. Nevertheless, combining existing data from different studies could lead to biased results due to the different strategies used

to generate data sets. In particular, while the DNA extraction method has been shown to produce a little impact on the microbial structure of samples with high microbial load (14), the diverse sequencing protocols could produce different results due to differences in sequence read length and different methodologies exploited to determine the nucleotide sequences. Accordingly, to achieve high resolution of the input data and avoid the above-mentioned bias, we focused only on metagenomic data sets obtained by Illumina sequencing platform.

In detail, shotgun metagenomic sequencing data of 207 fecal samples from 107 non-athletes and 100 athletes engaged in different types of sport (cyclist, rugby players, rower, runner, and marathon athletes) were collected from six different studies (9, 11, 15,–18) and submitted to a meta-analysis aimed at elucidating the microbial species composition (Table S1). After quality filtering and removal of reads mapping against the Homo sapiens genome, we obtained a collection of high-quality metagenomic samples with an average of 11,700,594 reads per sample (Table S1).

As previously suggested (7), the evaluation of the alpha-diversity, expressed as the species richness, showed statistically significant differences between the gut microbiomes of non-athletic individuals and athletes, with this latter showing a higher intestinal microbial biodiversity (average of 30 versus 34 species with relative abundance $> 0.05\%$, t test P-value < 0.05) (Table S2). Similarly, analysis of inter-individual variability through PCoA revealed statistically significant differences in the composition of fecal microbiota between athletes and non-athletes (PERMANOVA P-values < 0.05) regardless of ethnic-geographic location, gender, sport type, and study cohort (PERMANOVA P-values > 0.05),

reflecting the notion that exercise and exercise-related factors can shape the human gut microbial communities (Fig. S1a).

Taxonomic-based sample clustering and identification of the high prevalence clusters.

Hierarchical clustering (HCL) analysis was performed in combination with the Silhouette method (9), employing the species-level relative abundance data to capture recurrent different taxonomic profiles from metagenomic samples. This approach led to obtaining a statistically optimal number of 10 sample clusters based on their different bacterial composition, representing the community state types (CSTs), i.e., the recurring microbial patterns observed across the investigated cohort of individuals (Fig. 1a, Fig. S1b). Among these, three were identified as the most recurrent microbial profiles, referred to as high prevalence clusters (HPCs), covering individually at least 15% of the samples and collectively 73% of the subjects included in the meta-analysis (Fig. 1a, Table S3).

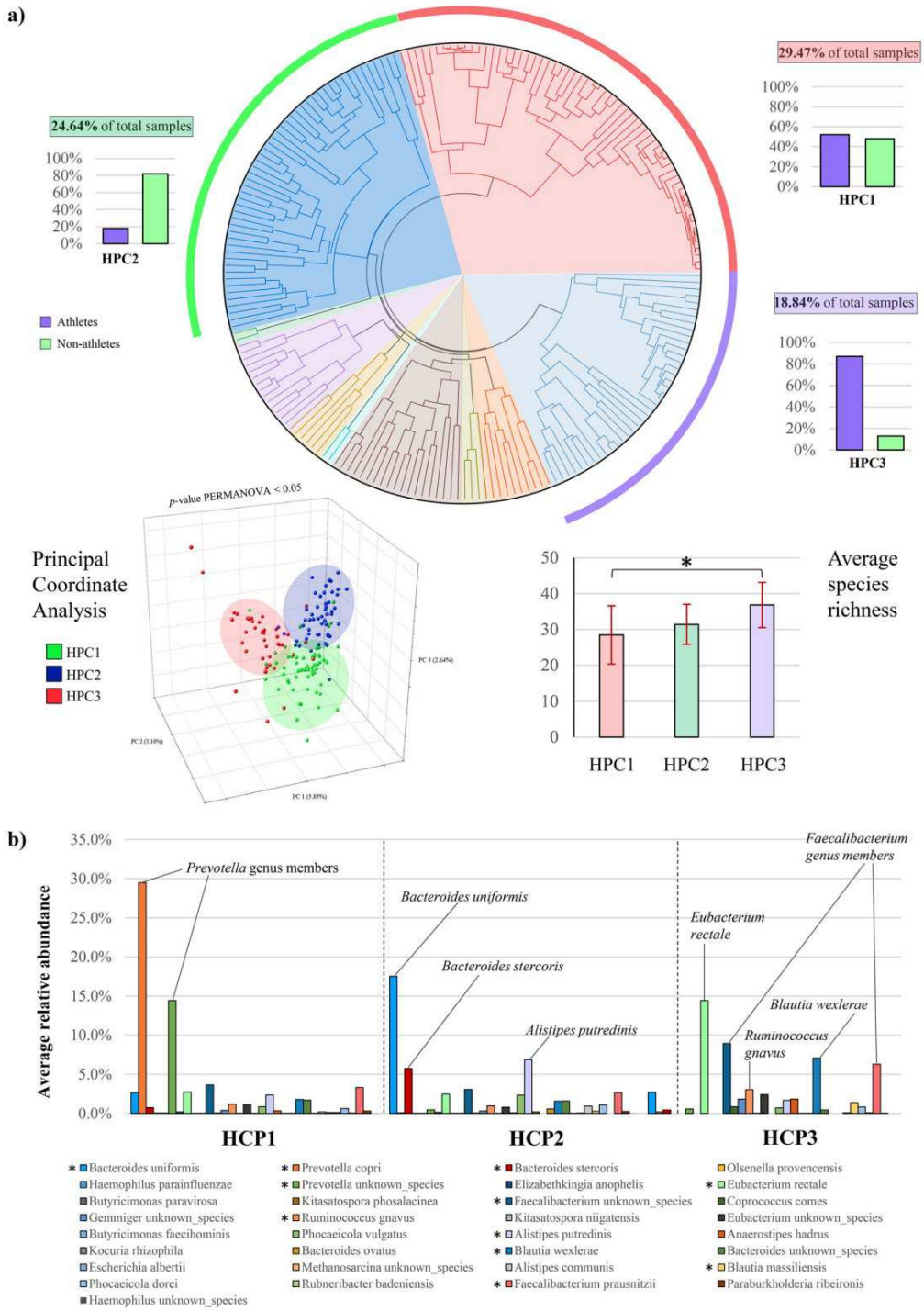

a)
24.64% of total samples
100%
80%
60%
40%
20%
0%
HPC2
29.47% of total samples
HPC1
18.84% of total samples
HPC3
Athletes
Non-athletes
p-value PERMANOVA < 0.05
Principal Coordinate Analysis
HPC1
HPC2
HPC3
PC 3 (2.64%)
PC 2 (3.10%)
PC 1 (5.85%)
50
40
30
20
10
0
*
HPC1
HPC2
HPC3
Average species richness
b)
Average relative abundance
35.0%
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
Prevotella genus members
Bacteroides uniformis
Bacteroides stercoris
Alistipes putredinis
Faecalibacterium genus members
Eubacterium rectale
Blautia wexlerae
Ruminococcus gnavus
HCP1
HCP2
HCP3
* Bacteroides uniformis
Haemophilus parainfluenzae
Butyricimonas paravirosa
Gemmiger unknown_species
Butyricimonas faecihominis
Kocuria rhizophila
Escherichia albertii
Phocaeicola dorei
Haemophilus unknown_species
* Prevotella copri
* Prevotella unknown_species
Kitasatospora phosalacinea
* Ruminococcus gnavus
Phocaeicola vulgatus
Bacteroides ovatus
Methanosarcina unknown_species
Rubneribacter badeniensis
* Bacteroides stercoris
Elizabethkingia anophelis
* Faecalibacterium unknown_species
Kitasatospora niigatensis
* Alistipes putredinis
* Blautia wexlerae
Alistipes communis
* Faecalibacterium prausnitzii
Olsenella provencensis
* Eubacterium rectale
Coprococcus comes
Eubacterium unknown_species
Anaerostipes hadrus
Bacteroides unknown_species
* Blautia massiliensis
Paraburkholderia ribeironis

Fig 1. Cluster analysis of the 100 athletes and 107 non-athletes subjects based on gut-associated microbial community composition. Panel a shows the circular HCL-based dendrogram resulting from metagenomic sample clustering that led to the definition of the three high prevalence cluster (HPCs). The proportions of metagenomic samples from athlete and non-athletic individuals in each HPC are reported through histograms outside the circle. Below, alpha- and beta-diversity analyses involving the three HPCs are depicted through a PCoA plot and a bar chart, respectively. In panel b, the microbial taxonomic composition is visualized through a bar chart showing the average relative abundance of each taxon at the species level. The main bacterial species showing statistically significant differences between HPCs are highlighted with asterisks on the chart legend.

Integration of the HCL analysis with the available metadata highlighted peculiar associations between HPCs and physical activity levels. In detail, while HPC1 showed a mixed composition (52% of athletes and 48% of control individuals), HPC2 encompassed 82% of non-athletes and HPC3 included 87% of athletes (Fig. 1a, Table S3). To note, after accounting for the study of origin, only 5% of the observed inter-samples variability was explained, demonstrating that geographic location and sample processing methods do not significantly impact on the microbial composition of the subjects included in HPC3 (Fig. S1c). Consistently, the subjects included in the above-mentioned HPCs showed statistically diverse gut microbiome composition, as evidenced by the principal coordinate analysis (PCoA) based on the microbial profiling data at the species level (Fig. 1a). Moreover, microbial biodiversity appears significantly higher in HPC3 (87% of athletes) compared with HPC1 (75% of non-athletes) (average species-richness of 36.8 versus 31.4) (Fig. 1a). As a result, at first glance, it seems that the gut microbiota of athletes is significantly diverse and more complex in

terms of taxonomic composition compared to those of subjects with a more sedentary lifestyle.

In particular, through the use of a polynomial linear model, which allows assessing the variability explained by each species (indicated as Adj. R-Square), we highlighted 33 taxa with a value greater than 0.15 (19), thus representing the bacterial species having the most impact in defining HPC structures (Table S3). In detail, these high-impact taxa covered from 45.86% to 68.80% of the three HPC bacterial compositions and highlighted clear connections between specific taxonomic patterns and the host's physical activity level, as discussed below (Table S3).

Dissection of the key microbial players of the gut microbiome of athletes and non-athletic individuals.

In order to catch the association between physical activity and specific taxa, we focused on the 33 microbial taxa individuated above as responsible for the main compositional differences between the three HPCs.

In particular, HPC1, composed of 52% of athletes and 48% of non-athletes, was distinct in having high relative abundances of *Prevotella* genus members (average relative abundance of 43.9%) (Fig. 1b, Table S3), which are considered a common commensal microorganism often associated with high dietary fiber intakes (20). In contrast, HPC2, composed of 82% of samples from non-athletic individuals, was defined by the presence of *Bacteroides* members, including *Bacteroides uniformis* with an average relative abundance of 17.5% (Fig. 1b, Table S3), as expected from healthy subjects (21). Indeed, the Bacteroides taxon is well-known to represent a large portion of the dominant healthy human gut microbiota, previously reported to characterize one of the three renowned human

enterotypes (22). Nevertheless, based on HPC2 composition, a non-athletic lifestyle was associated with increased *Alistipes putredinis* abundance (average relative abundance of 5.9%) compared with individuals with high physical activity, i.e., HPC3 (Fig. 1b, Table S3). This taxon is a member of a relatively recent genus taxonomically closely related to the Bacteroidetes phylum (23), whose role in the gut ecosystem is controversial (24). However, previous studies have suggested an association between *Alistipes* and inflammation and disease, including cardiovascular disease and colorectal cancer (25, 26).

Of note, HPC3, composed for the 87% of athletes, is characterized by members of *Faecalibacterium* genus, along with *Eubacterium rectale* and *Blautia wexlerae*, with average relative abundances of 15.2%, 14.4%, and 7.1%, respectively, thus resulting significantly higher than those of non-athletic individuals (P-values < 0.05) (Fig. 1b, Table S3). Interestingly, *F. prausnitzii*, *E. rectale*, and members of the *Blautia* genus have been linked with beneficial effects in various clinical conditions, including inflammatory bowel diseases, metabolic syndromes, and colorectal cancer (27,–29). Moreover, these taxa have been reported to be responsible for butyrate production (30,–32), contributing not only to intestinal anti-inflammatory effects but also to host energy metabolism through de novo synthesis of glucose and lipids, which are primary sources of energy for the host organism (33, 34).

Remarkably, these findings revealed clear structural differences between the gut microbiota of the athletes and that of subjects with no physical activity, suggesting the importance of athlete gut-associated microorganisms both as supporters of the gut homeostasis as well as a source of compounds that can increase energy harvest, thus possibly improving athlete performances. However, the limited availability of precise information regarding the individual nutrition

regimen did not allow further investigation of the correlation between diet and gut microbiota composition. Thus, future studies will need to collect as a wide range of metadata as possible, including dietetic regimes, that could be essential to understanding how exercise and exercise-associated factors affect the gut microbiota-host interactions in athletes.

Analysis of the interaction networks sustaining the gut microbial community of athletes and non-athletes.

In order to explore the intricate interaction network of the multispecies community constituting the three HPCs, we performed a microbial co-occurrence analysis aimed at highlighting the degree of displacement (negative links) or coexistence (positive links) between species (Table S4). Correlation data were represented by a network of nodes (microbial species) linked in pairs by green edges when the relationships were positive or red edges when they were negative. Furthermore, modularity clusters (MCs) analysis allowed to detect community (sub)structures in networks, i.e., groups of taxa highly interconnected (Fig. 2, Fig. 3). Interestingly, the comparison between the network describing the gut-associated microbial community from athletes and non-athletes revealed a marked difference in the number of statistically significant interactions among taxa (positive and negative links) (Fig. 2). In particular, the microbial network of athletes showed 328 statistically significant associations, of which 62% were positive, in contrast to a total of 223 found gut microbiota members of non-athletic individuals (Table S4). Generally, compared with relatively simple networks, complex interconnected networks have a higher nutritional interaction among community members, such as cross-feeding of essential small molecules,

resulting in a more stable microbial consortium with improved resilience to ecosystem disturbances (35). In addition, among the taxa with a prominent role in athlete's gut microbiota structure, we found species belonging to *Faecalibacterium*, *Eubacterium*, *Ruminococcus*, and *Blautia* genera that are thought to promote intestinal barrier integrity and prevent inflammation (36). Accordingly, these results suggested that the microbial community of athletes exhibits improved stability compared with the gut microbiome of non-athletic individuals, pointing to the importance of microbial synergism among health-promoting species in sustaining the exercise-induced microbiome changes.

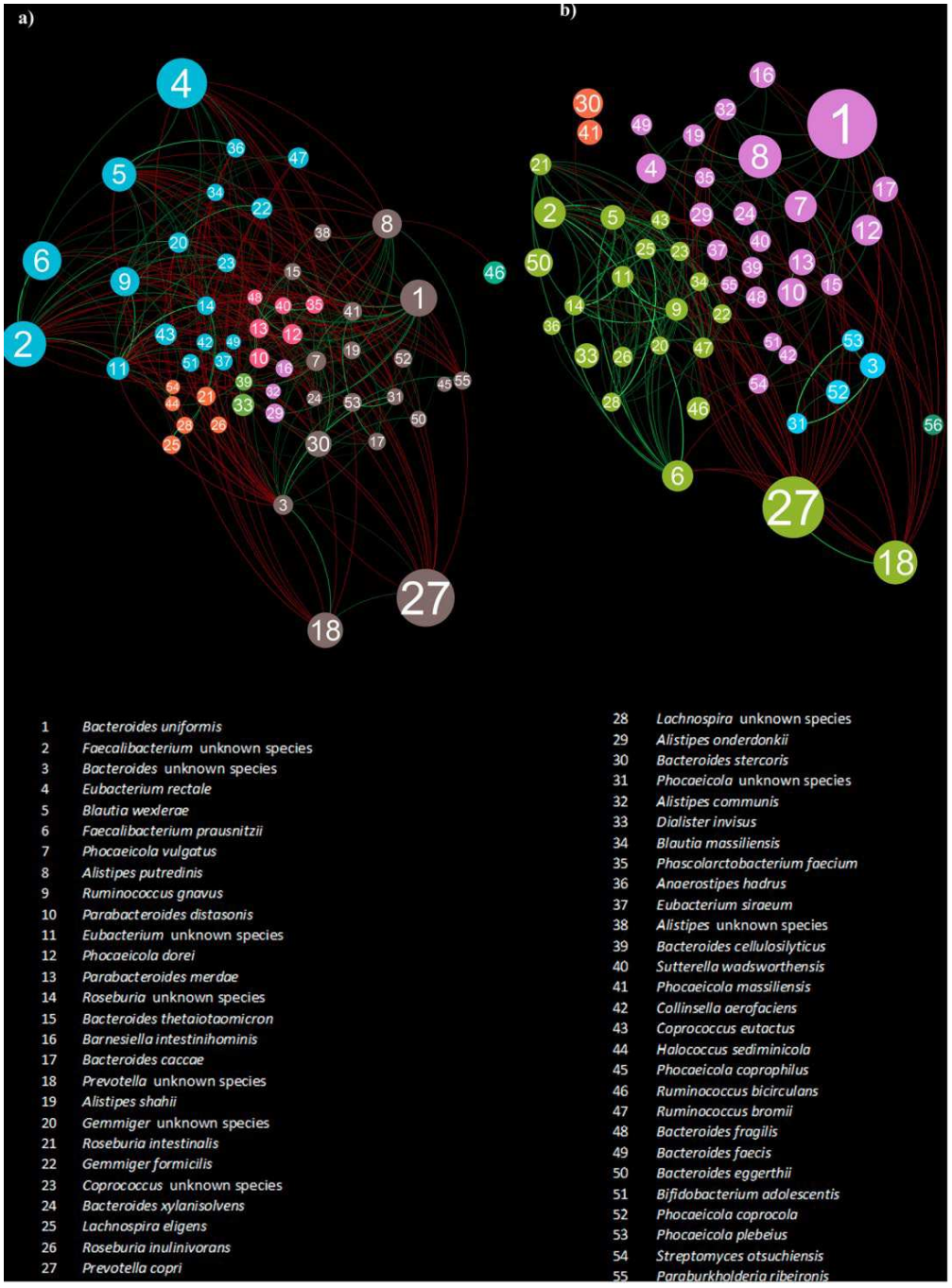


Fig 2. Interaction network supporting the structure of the gut microbial consortia in athletes and non-athletes. Panel a reports the interaction network of athlete gut microbiota, and panel b depicts the interaction network of the fecal microbial community of non-athletic individuals. In the force-

driven networks, nodes represent bacterial taxa, and covariance values were used to construct the edges. Red edges correspond to negative correlations, while green edges represent positive associations. The node size is proportional to the relative average abundance of each taxon.

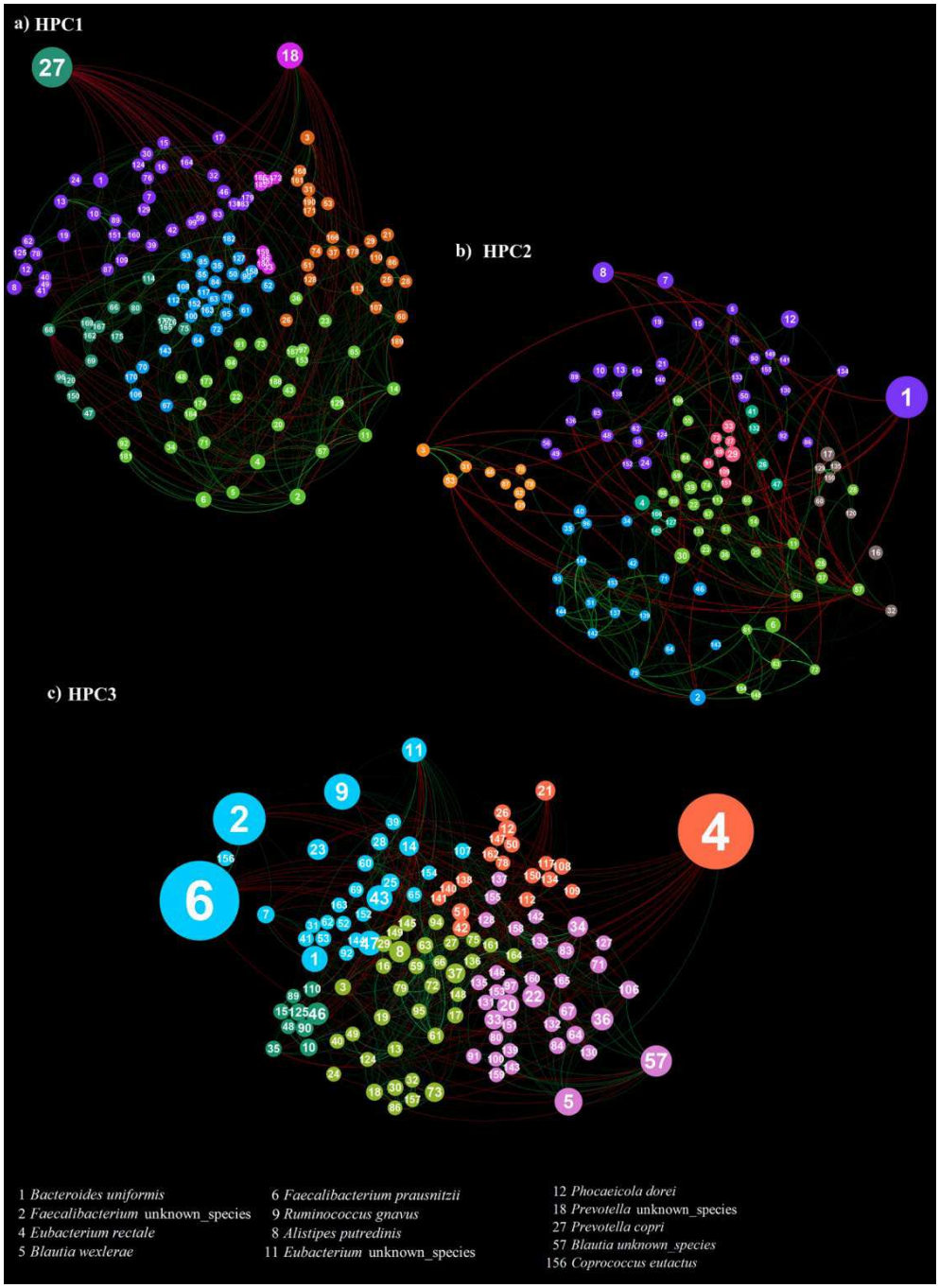


Fig 3. Co-occurrence network characterizing the three HPCs. The networks visualize the covariance relationships between the microbial taxa composing HPC1 (panel a), HPC2 (panel b), and HPC3 (panel c). HPC1 encompasses 52% of athletes and 48% of non-athletic subjects, HPC2 contains 82% of non-athletes, and HPC3 contains 87% of athletes. The complete one-to-one correspondence between node labels and microbial taxa is available in Table S5.

Co-occurrence network analyses of HPCs1 to 3.

Focusing on individual HPC-derived networks, microbial correlation analysis of HPC1, which showed a mixed composition of non-athletic and athletic individuals, it is worth mentioning that members of *Prevotella* genus (node 27 and 18), such as *Prevotella copri* (node 27), tend to dominate their intestinal ecological niche. In addition, this taxon negatively correlated with other typical key members of the healthy gut-associated microbial communities, including *B. uniformis*, *Ruminococcus gnavus*, and members of *Faecalibacterium* genus (Fig. 3a, Table S4). Simultaneously, a dense and intricate network of positive associations between minority players (proportion of 94% of the total network interactions) seems to sustain the microbial community structure of HPC1.

Conversely, the HPC2, which covers mainly non-athletic subjects, appeared to be driven by five related keystone taxa, belonging to *Bacteroides* (nodes 1 and 30), *Phocaeicola* (nodes 12 and 41), and *Alistipes* (node 8) genera (Fig. 3b, Table S4). In particular, these taxa were engaged in negative correlations mainly with potentially anti-inflammatory, butyrate-producing bacteria from the genera *Ruminoccoccus*, *Faecalibacterium*, and *Blautia* (28, 37), thus revealing a possible negative impact of a sedentary or low physical activity lifestyle on health-associated commensal bacteria. However, a small-scale subnetwork (light

blue) comprising well-known commensals of the healthy human gut microbiota, such as *Bifidobacterium longum*, *Bifidobacterium adolescentis*, and *Collinsella aerofaciens* (Fig. 3b, Table S4), despite their low relative abundance in non-athletic subjects (<1%), seem to play a pivotal role in establishing positive correlations with other minor microbial players, regulating a large part of the microbial consortium characterizing healthy non-athletic individuals.

Interestingly, interaction networks describing the gut-associated microbial community of athletes, i.e., HPC3 (Fig. 3c, Table S4), showed the highest number of species that, being involved in conspicuous biotic interactions, seem to influence the whole-community dynamics of the athlete gut microbiota. Indeed, as previously mentioned, health-associated species, i.e., *Faecalibacterium prausnitzii* (node 6), *Blautia wexlerae* (node 5), and *Eubacterium rectale* (node 4), along with *Ruminococcus gnavus* (node 9), act as keystone taxa in HPC3, exerting considerable control on the entire community structure (Fig. 3c, Table S4). In addition, these taxa are involved in strong positive associations (Spearman correlation coefficient value > 0.5) with members of the *Coprococcus* and *Roseburia* genera that, being part of commensal bacteria producing SCFAs, primarily butyrate, exert a positive influence on intestinal barrier maintenance, colonic motility, and anti-inflammatory processes (38,–40). Besides, additional low-abundance members appear to have significant effects on the intestinal niche, reflecting the existence of a complex and solid ecosystem. As a result, removing a few species likely does not lead to a dramatic shift in the composition. Taken together, these findings support the notion that exercise can affect the gut microbiota composition, inducing qualitative and quantitative changes that may confer beneficial effects to the host and possibly to athletic performance.

Conclusion

Accumulating evidence has suggested a bidirectional association between physical activity and the composition of the microbial communities inhabiting the human intestinal environment (41). Indeed, differences in the gut microbiota composition have been observed between athletes and non-athletes, with this latter showing an increased abundance of short-chain fatty acids (SCFAs)-producing bacterial species (8, 42). In turn, the gut microbiota is thought to play a significant role in amino acid and carbohydrate host metabolism, likely indirectly influencing athlete health, training, sports performance, and post-exercise recovery (41, 43).

In this framework, a metagenomic analysis was performed by exploiting publicly available shotgun metagenomic data sets with the aim to provide insights into the gut-associated microbial community structure in athletes. In particular, a collection of 100 metagenomic samples from athletes and 107 from healthy non-athletic individuals allowed us to identify three high prevalence clusters (HPC1 to 3), i.e., recurring patterns of microbial composition. Interestingly, the gut microbiome of athletes (HPC3) showed higher biodiversity with an increased abundance of gut-associated health-promoting bacterial species compared to non-athletes. In particular, SCFAs-producing species such as *F. prausnitzii*, *E. rectale*, *B. wexlerae*, and *R. gnavus*, were associated with athlete physical activity, revealing their possible contribution to the host health, regulating inflammation and immune system, as well as athlete's energy acquisition and sport performances. Moreover, an intricate and solid network of biotic interactions sustained by seven health-promoting key species and a range of concurrent low-abundance taxa seems to characterize the microbial community

of athletes. In contrast, a less clustered and less inter-connected network was obtained from non-athletic subjects. Based on these findings, it appears that exercise induces gut microbiota changes resulting in an increased abundance of bacteria with potential health benefits, such as SCFAs producers, cooperating in complex, interconnected microbial communities, with possible positive implications on sports performance. Future detailed functional analysis addressing the metabolic capability of the gut microbiota will aid in elucidating the connection between microbial-derived metabolites and athletic versus non-athletic lifestyle.

Material and Methods

Metagenomic sample collection.

With the aim to explore the differences in the gut microbiome composition between athletes and non-athletic individuals, we retrieved all the publicly available shotgun metagenomic raw data (fastq) from the National Center of Biotechnology Information (NCBI) Sequence Read Archive (SRA) database. Accordingly, to safeguard consistency and equivalence across metagenomic samples from different studies, we selected only those produced through Illumina sequencing method. As a result, we collected 207 shotgun metagenomics samples from six different studies (PRJEB15388, PRJEB28338, PRJEB32794, PRJNA472785, PRJNA305507, PRJEB20054), of which 100 corresponded to athlete gut microbiomes, and 107 were from healthy non-athletes (Table S1). In addition, the respective metadata regarding health status, training type, exercise intensity level, and diet were also collected (Table S1).

Metagenomics data processing and taxonomic profiling.

The fastq raw data obtained from publicly repositories were submitted to quality filtering to remove sequence reads with low quality scores (<25). Subsequently, removal of reads mapping on the hg19 human reference genome was performed to exclude host DNA. This process allowed to achieve an average of 11,700,594 ± 9,886,096 reads per sample that were submitted to downstream analyses. The retained reads were subjected to taxonomic classification using METAnnotatorX2 bioinformatics platform (44), which performs MegaBLAST local alignment of reads (45) to the curated non-redundant sequence database of genomes retrieved from NCBI servers.

For each metagenomic sample, taxonomical biodiversity, i.e., species richness, was calculated as the number of gut-associated bacterial taxa whose sequenced reads had a relative abundance greater than 0.5%. Similarities between samples (beta-diversity) were calculated by Bray-Curtis dissimilarity based on species abundance. The range of similarities is calculated between values 0 and 1. PCoA representation of beta-diversity was performed using ORIGIN 2021 (https://www.originlab.com/2021). In the PCoA each dot represented a sample, distributed in tridimensional space according to its own bacterial composition.

The hierarchical clustering (HCL) of samples was achieved employing bacterial composition at the species level and was calculated through TMeV 4.8.1 software using Pearson correlation as a distance metric based on species-level information. The data obtained was represented by a dendrogram.

Microbial co-occurrence and network analyses.

Covariance analysis involving the 332 bacterial species obtained by taxonomic profiling of the 207 metagenomic fecal samples was realized employing Kendall's tau rank covariance analysis (46). Using software Gephi (https://gephi.org/), the obtained correlation coefficients were exploited to build a force-driven network, whose nodes represent bacterial species, and edges define their relationships. The node size is related to the number of interactions of a specific microbial taxon, i.e., the node degree, while the edge color shows the type of interaction, i.e., positive (green) or negative (red).

Statistical analysis.

ORIGIN 2021 (https://www.originlab.com/2021) and SPSS software (www.ibm.com/software/it/analytics/spss/) were used to compute statistical analyses. PERMANOVA analyses were performed using 1,000 permutations to assess p-values for differences among populations in PCoA analyses. Furthermore, bacterial abundance differences were tested by t-test analysis.

Acknowledgments

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

Reference

1. Rowland I, Gibson G, Heinken A, Scott K, Swann J, Thiele I, Tuohy K. 2018. Gut microbiota functions: metabolism of nutrients and other food components. Eur J Nutr 57:1–24. doi: 10.1007/s00394-017-1445-8.
2. Yoo JY, Groer M, Dutra SVO, Sarkar A, McSkimming DI. 2020. Gut microbiota and immune system interactions. Microorganisms 8:1587–1522. doi: 10.3390/microorganisms8101587.
3. Hakansson A, Molin G. 2011. Gut microbiota and inflammation. Nutrients 3:637–682. doi: 10.3390/nu3060637.
4. Hasan N, Yang H. 2019. Factors affecting the composition of the gut microbiota, and its modulation. PeerJ 7. doi: 10.7717/peerj.7502.
5. Hughes RL, Holscher HD. 2021. Fueling gut microbes: a review of the interaction between diet, exercise, and the gut microbiota in athletes. Adv Nutr 12:2190–2215. doi: 10.1093/advances/nmab077.
6. Piercy KL, Troiano RP, Ballard RM, Carlson SA, Fulton JE, Galuska DA, George SM, Olson RD. 2018. The physical activity guidelines for Americans. JAMA 320:2020–2028. doi: 10.1001/jama.2018.14854.
7. Clarke SF, Murphy EF, O'Sullivan O, Lucey AJ, Humphreys M, Hogan A, Hayes P, O'Reilly M, Jeffery IB, Wood-Martin R, Kerins DM, Quigley E, Ross RP, O'Toole PW, Molloy MG, Falvey E, Shanahan F, Cotter PD. 2014. Exercise and associated dietary extremes impact on gut microbial diversity. Gut 63:1913–1920. doi: 10.1136/gutjnl-2013-306541.
8. Hughes RL. 2020. A review of the role of the gut microbiome in personalized sports nutrition. Front Nutr 6. doi: 10.3389/fnut.2019.00191.
9. Scheiman J, Luber JM, Chavkin TA, MacDonald T, Tung A, Pham L-D, Wibowo MC, Wurth RC, Punthambaker S, Tierney BT, Yang Z, Hattab MW, Avila-Pacheco J, Clish CB, Lessard S, Church GM, Kostic AD. 2019. Meta-omics analysis of elite athletes identifies a performance-enhancing microbe that functions via lactate metabolism. Nat Med 25:1104–1109. doi: 10.1038/s41591-019-0485-4.
10. Turpin-Nolan SM, Joyner MJ, Febbraio MA. 2019. Can microbes increase exercise performance in athletes? Nat Rev Endocrinol 15:629–630. doi: 10.1038/s41574-019-

0250-2.

11. Petersen LM, Bautista EJ, Nguyen H, Hanson BM, Chen L, Lek SH, Sodergren E, Weinstock GM. 2017. Community characteristics of the gut microbiomes of competitive cyclists. Microbiome 5:98. doi: 10.1186/s40168-017-0320-4.
12. Blomstrand E, Eliasson J, Karlsson HKR, Köhnke R. 2006. Branched-chain amino acids activate key enzymes in protein synthesis after physical exercise. The J Nutrition 136:269S–273S. doi: 10.1093/jn/136.1.269S.
13. Spriet LL. 2019. Performance Nutrition for Athletes. Sports Medicine (Auckland, NZ) 49. https://pubmed.ncbi.nlm.nih.gov/30671901/. [PMC free article] [PubMed] [Google Scholar]
14. Sui H-y, Weil AA, Nuwagira E, Qadri F, Ryan ET, Mezzari MP, Phipatanakul W, Lai PS. 2020. Impact of DNA extraction method on variation in human and built environment microbial community and functional profiles assessed by shotgun metagenomics sequencing. Front Microbiol 11. doi: 10.3389/fmicb.2020.00953.
15. Barton W, Penney NC, Cronin O, Garcia-Perez I, Molloy MG, Holmes E, Shanahan F, Cotter PD, O'Sullivan O. 2018. The microbiome of professional athletes differs from that of more sedentary subjects in composition and particularly at the functional metabolic level. Gut 67:625–633. https://pubmed.ncbi.nlm.nih.gov/28360096/. [PubMed] [Google Scholar]
16. O'Donovan CM, Connor B, Madigan SM, Cotter PD, O'Sullivan O. 2020. Instances of altered gut microbiomes among Irish cricketers over periods of travel in the lead up to the 2016 World Cup: a sequencing analysis. Travel Medicine and Infectious Dis 35:101553. doi: 10.1016/j.tmaid.2020.101553.
17. O'Donovan CM, Madigan SM, Garcia-Perez I, Rankin A, O'Sullivan O, Cotter PD. 2020. Distinct microbiome composition and metabolome exists across subgroups of elite Irish athletes. J Sci Med Sport 23:63–68. doi: 10.1016/j.jsams.2019.08.290.
18. Cronin O, Barton W, Skuse P, Penney NC, Garcia-Perez I, Murphy EF, Woods T, Nugent H, Fanning A, Melgar S, Falvey EC, Holmes E, Cotter PD, O'Sullivan O, Molloy MG, Shanahan F. 2018. A prospective metagenomic and metabolomic analysis of the impact of exercise and/or whey protein supplementation on the gut microbiome of sedentary

adults. mSystems 3. doi: 10.1128/mSystems.00044-18.

19. Kelsey CM, Prescott S, McCulloch JA, Trinchieri G, Valladares TL, Dreisbach C, Alhusen J, Grossmann T. 2021. Gut microbiota composition is associated with newborn functional brain connectivity and behavioral temperament. Brain Behav Immun 91:472–486. doi: 10.1016/j.bbi.2020.11.003.
20. Kovatcheva-Datchary P, Nilsson A, Akrami R, Lee YS, De Vadder F, Arora T, Hallen A, Martens E, Björck I, Bäckhed F. 2015. Dietary fiber-induced improvement in glucose metabolism is associated with increased abundance of Prevotella. Cell Metab 22:971–982. doi: 10.1016/j.cmet.2015.10.001.
21. Zafar H, Saier MH. 2021. Gut Bacteroides species in health and disease. Gut Microbes 13:1–20. https://pubmed.ncbi.nlm.nih.gov/33535896/. [PMC free article] [PubMed] [Google Scholar]
22. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto J-M, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, de Vos WM, Brunak S, Doré J, Weissenbach J, Ehrlich SD, Bork P, MetaHIT Consortium (additional members). 2011. Enterotypes of the human gut microbiome. Nature 473:174–180. doi: 10.1038/nature09944.
23. Rautio M, Eerola E, Väisänen-Tunkelrott M-L, Molitoris D, Lawson P, Collins MD, Jousimies-Somer H. 2003. Reclassification of Bacteroides putredinis (Weinberg et al., 1937) in a new genus Alistipes gen. nov., as Alistipes putredinis comb. nov., and description of Alistipes finegoldii sp. nov., from human sources. Syst Appl Microbiol 26:182–188. doi: 10.1078/072320203322346029.
24. Parker BJ, Wearsch PA, Veloo ACM, Rodriguez-Palacios A. 2020. The genus alistipes: gut bacteria with emerging implications to inflammation, cancer, and mental health. Front Immunol 11:906. doi: 10.3389/fimmu.2020.00906.
25. Jie Z, Xia H, Zhong S-L, Feng Q, Li S, Liang S, Zhong H, Liu Z, Gao Y, Zhao H, Zhang D, Su Z, Fang Z, Lan Z, Li J, Xiao L, Li J, Li R, Li X, Li F, Ren H, Huang Y, Peng Y, Li G,

Wen B, Dong B, Chen J-Y, Geng Q-S, Zhang Z-W, Yang H, Wang J, Wang J, Zhang X, Madsen L, Brix S, Ning G, Xu X, Liu X, Hou Y, Jia H, He K, Kristiansen K. 2017. The gut microbiome in atherosclerotic cardiovascular disease. Nat Commun 8. doi: 10.1038/s41467-017-00900-1.

26. Moschen AR, Gerner RR, Wang J, Klepsch V, Adolph TE, Reider SJ, Hackl H, Pfister A, Schilling J, Moser PL, Kempster SL, Swidsinski A, Orth Höller D, Weiss G, Baines JF, Kaser A, Tilg H. 2016. Lipocalin 2 protects from inflammation and tumorigenesis associated with gut microbiota alterations. Cell Host Microbe 19:455–469. doi: 10.1016/j.chom.2016.03.007.
27. Mukherjee A, Lordan C, Ross RP, Cotter PD. 2020. Gut microbes from the phylogenetically diverse genus Eubacterium and their various contributions to gut health. Gut Microbes 12:1802866. doi: 10.1080/19490976.2020.1802866.
28. Liu X, Mao B, Gu J, Wu J, Cui S, Wang G, Zhao J, Zhang H, Chen W. 2021. Blautia-a new functional genus with potential probiotic properties? Gut Microbes 13:1–21. https://pubmed.ncbi.nlm.nih.gov/33525961/. [PMC free article] [PubMed] [Google Scholar]
29. Ferreira-Halder CV, Faria A. V d S, Andrade SS. 2017. Action and function of Faecalibacterium prausnitzii in health and disease. Best Pract Res Clin Gastroenterol 31:643–648. doi: 10.1016/j.bpg.2017.09.011.
30. Nilsen M, Madelen Saunders C, Leena Angell I, Arntzen MØ, Lødrup Carlsen KC, Carlsen K-H, Haugen G, Heldal Hagen L, Carlsen MH, Hedlin G, Monceyron Jonassen C, Nordlund B, Maria Rehbinder E, Skjerven HO, Snipen L, Cathrine Staff A, Vettukattil R, Rudi K. 2020. Butyrate levels in the transition from an infant- to an adult-like gut microbiota correlate with bacterial networks associated with Eubacterium rectale and Ruminococcus gnavus. Genes 11:1245–1215. doi: 10.3390/genes11111245.
31. Morrison DJ, Preston T. 2016. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. Gut Microbes 7:189–200. doi: 10.1080/19490976.2015.1134082.
32. Vacca M, Celano G, Calabrese FM, Portincasa P, Gobbetti M, de Angelis M. 2020. The controversial role of human gut lachnospiraceae. Microorganisms [Internet] 8:573. doi:

10.3390/microorganisms8040573.

33. den Besten G, van Eunen K, Groen AK, Venema K, Reijngoud DJ, Bakker BM. 2013. The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. J Lipid Res 54:2325–2340. doi: 10.1194/jlr.R036012.

34. den Besten G, Lange K, Havinga R, van Dijk TH, Gerding A, van Eunen K. 2013. Gut-derived short-chain fatty acids are vividly assimilated into host carbohydrates and lipids. American J Physiology Gastrointestinal and Liver Physiology 305. doi: 10.1152/ajpgi.00265.2013.

35. Wagg C, Schlaeppi K, Banerjee S, Kuramae EE, van der Heijden MGA. 2019. Fungal-bacterial diversity and microbiome complexity predict ecosystem functioning. Nat Commun 10. doi: 10.1038/s41467-019-12798-y.

36. Lordan C, Thapa D, Ross RP, Cotter PD. 2020. Potential for enriching next-generation health-promoting gut bacteria through prebiotics and other dietary components. Gut Microbes 11:1–20. doi: 10.1080/19490976.2019.1613124.

37. Takahashi K, Nishida A, Fujimoto T, Fujii M, Shioya M, Imaeda H, Inatomi O, Bamba S, Andoh A, Sugimoto M. 2016. Reduced abundance of butyrate-producing bacteria species in the fecal microbial community in Crohn's disease. Digestion 93:59–65. doi: 10.1159/000441768.

38. Nie K, Ma K, Luo W, Shen Z, Yang Z, Xiao M, Tong T, Yang Y, Wang X. 2021. Roseburia intestinalis: a beneficial gut organism from the discoveries in genus and species. Front Cell Infect Microbiol 11:757718. doi: 10.3389/fcimb.2021.757718.

39. Valles-Colomer M, Falony G, Darzi Y, Tigchelaar EF, Wang J, Tito RY, Schiweck C, Kurilshikov A, Joossens M, Wijmenga C, Claes S, Van Oudenhove L, Zhernakova A, Vieira-Silva S, Raes J. 2019. The neuroactive potential of the human gut microbiota in quality of life and depression. Nat Microbiol 4:623–632. doi: 10.1038/s41564-018-0337-x.

40. Canani RB, di Costanzo M, Leone L, Pedata M, Meli R, Calignano A. 2011. Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. World J Gastroenterol 17:1519–1528. doi: 10.3748/wjg.v17.i12.1519.

41. Aya V, Flórez A, Perez L, Ramírez JD. 2021. Association between physical activity and

changes in intestinal microbiota composition: A systematic review. PLoS One 16:e0247039. doi: 10.1371/journal.pone.0247039.

42. Mohr AE, Jäger R, Carpenter KC, Kerksick CM, Purpura M, Townsend JR, West NP, Black K, Gleeson M, Pyne DB, Wells SD, Arent SM, Kreider RB, Campbell BI, Bannock L, Scheiman J, Wissent CJ, Pane M, Kalman DS, Pugh JN, Ortega-Santos CP, ter Haar JA, Arciero PJ, Antonio J. 2020. The athletic gut microbiota. J Int Soc Sports Nutr 17. doi: 10.1186/s12970-020-00353-w.
43. Koh A, de Vadder F, Kovatcheva-Datchary P, Bäckhed F. 2016. From dietary fiber to host physiology: short-chain fatty acids as key bacterial metabolites. Cell 165:1332–1345. doi: 10.1016/j.cell.2016.05.041.
44. Milani C, Lugli GA, Fontana F, Mancabelli L, Alessandri G, Longhi G, Anzalone R, Viappiani A, Turroni F, van Sinderen D, Ventura M. 2021. METAnnotatorX2: a comprehensive tool for deep and shallow metagenomic data set analyses. mSystems 6. doi: 10.1128/mSystems.00583-21.
45. Chen Y, Ye W, Zhang Y, Xu Y. 2015. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res 43:7762–7768. doi: 10.1093/nar/gkv784.
46. Liu X, Ning J, Cheng Y, Huang X, Li R. 2019. A flexible and robust method for assessing conditional association and conditional concordance. Stat Med 38:3656–3668. doi: 10.1002/sim.8202.

Chapter 4

The human gut microbiome of athletes: metagenomic and metabolic insights

Federico Fontana, Giulia Longhi, Chiara Tarracchini, Leonardo Mancabelli, Gabriele Andrea Lugli, Giulia Alessandri, Francesca Turroni, Christian Milani* and Marco Ventura*

The results of this chapter were published in Microbiome, 2023 Feb 14; https://doi.org/10.1186/s40168-023-01470-9.

* These authors contributed equally.

Abstract

Background

The correlation between the physical performance of athletes and their gut microbiota has become of growing interest in the past years, since new evidences have emerged regarding the importance of the gut microbiota as a main driver of the health status of athletes. In addition, it has been postulated that the metabolic activity of the microbial population harbored by the large intestine of athletes might influence their physical performances. Here, we analyzed 418 publicly available shotgun metagenomics datasets obtained from fecal samples of healthy athletes and healthy sedentary adults.

Results

This study evidenced how agonistic physical activity and related lifestyle can be associated with the modulation of the gut microbiota composition, inducing modifications of the taxonomic profiles with an enhancement of gut microbes able to produce short-fatty acid (SCFAs). In addition, our analyses revealed a correlation between specific bacterial species and high impact biological synthases (HIBSs) responsible for the generation of a range of microbially driven compounds such vitamin B12, amino acidic derivatives, and other molecules linked to cardiovascular and age-related health-risk reduction.

Conclusions

Notably, our findings show how subsist an association between competitive athletes, and modulation of the gut microbiota, and how this modulation is reflected in the potential production of microbial metabolites that can lead to beneficial effects on human physical performance and health conditions.

Introduction

In recent years, the increasing interest on the gut microbiota revealed how its relationship with the host is not limited to the intestinal environment but affects the entire human body across all the life stages, from birth to elderly [1, 2]. Stress and unbalanced diets are just two of the key drivers modulating the gut microbiota composition, shifting it towards a dysbiosis state, with potential negative impacts on systemic health [3–9]. On the contrary, a gut microbiota in homeostatic equilibrium is considered stable and able to maximize the beneficial interactions of the various members of the microbiota with the host, showing the capability of resisting external and internal influences [10].

While diet is one of the most impactful factors shaping the gut microbiota composition, physical activity can also modulate the gut microbiota through many mechanisms, such as the increased release of hormones and the redirection of blood from the gut to the skeletal muscles [11, 12]. In detail, the type of training, intensity, and duration of the physical activities impact the gut microbial population, ultimately altering its enzymatic potential responsible for systemic effects on the human host [12]. For example, studies concerning athletes have shown that they may be more susceptible to developing Inflammatory Bowel Diseases (IBD) [11, 13–16]. However, healthy athletes showed an increase in the production of short-fatty acid (SCFAs) for a greater energy intake, thereby contributing to host global metabolic efficiency [11, 17–19].

Remarkably, microbial SCFAs producers have been reported to generally possess a vast repertoire of metabolic pathways, not limited only to energy-related metabolism (i.e., short-chain fatty acid synthesis) but also including enzymes for

amino acid and vitamin metabolism as well as for the synthesis of other by-products [20, 21].

However, despite the great scientific interest of this topic, the available scientific literature mainly focus on a limited range of well-known microbial taxa involved in the production of few metabolites, such as lactic acid and short-chain fatty acids. This is in contrast with the vast number of the microbial metabolic pathways encompassed by the gut microbiomes and therefore the high number of the potentially microbial produced health-active metabolites [3]. Thus, little is still known regarding the physiological mechanisms involving resident bacteria modulated by physical activity and their impacts on the host in terms of physical performances and systemic health. For this reason, it is becoming pivotal to gain insights into this intricate network of metabolic host-microbes' interactions by analyzing in detail the gut microbiota composition in correlation with its genetic potential.

To delve into this intriguing area, in this study, we correlated physical activity metadata with taxonomical and microbial metabolic profiles of the gut microbiomes involving 185 athletes, 69 moderate athlete, and 166 controls (sedentary), using an in silico approach based on statistical analysis and correlations as well as hierarchical clustering and an optimized pipeline for metagenomic analysis.

Results and discussion

Metagenomic data selection and meta-analysis.

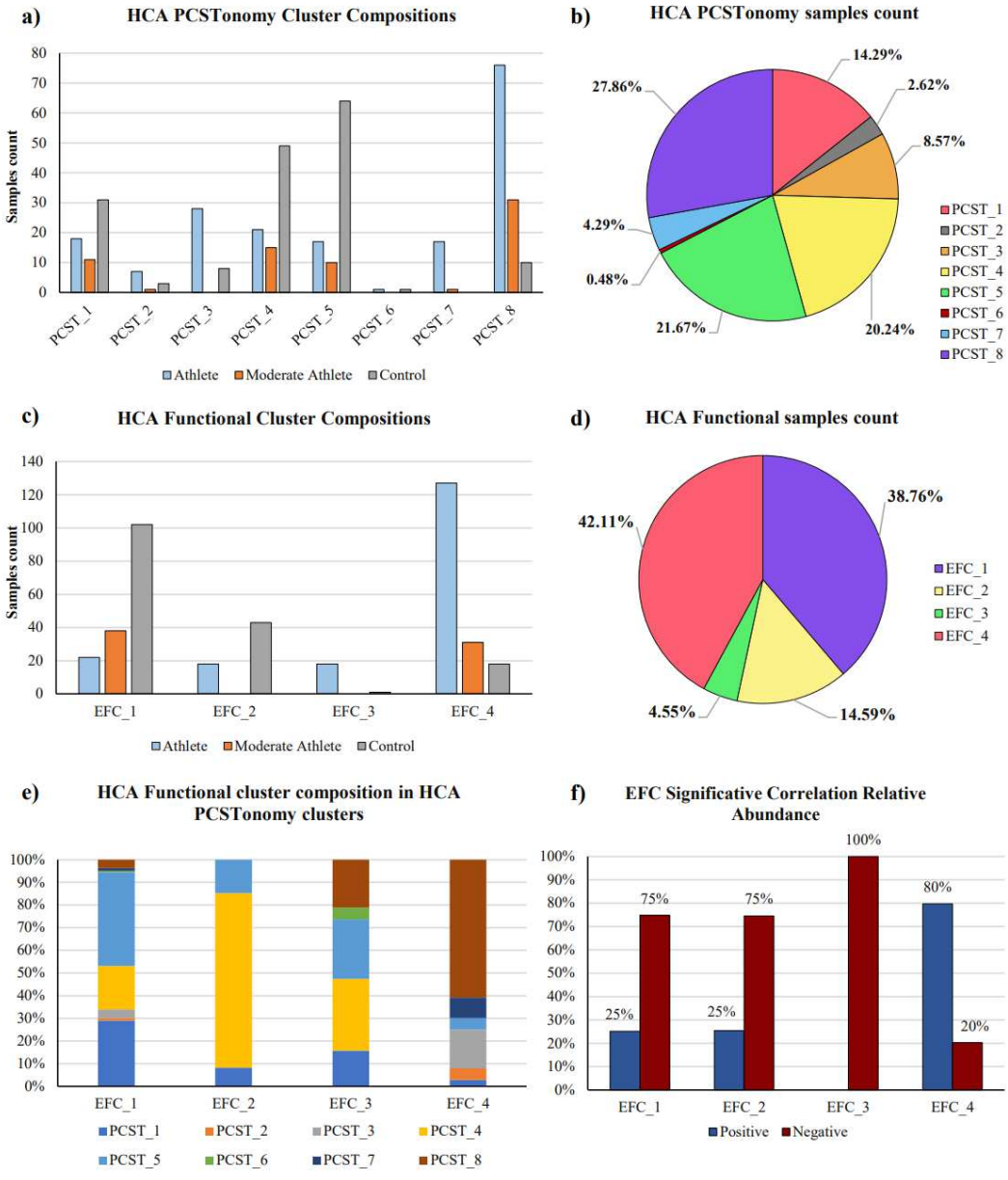
In order to determine how physical activity can be associated to the modification in the composition of the gut microbiota and vice versa, the NCBI repository was screened for shotgun metagenomic samples related to the gut microbiota of professional athletes. Specifically, we used athletes' metagenomics samples from multiple Bioprojects obtained from the same sequencing technology to avoid sampling-related bias. This screening resulted in the selection of a total of 185 metagenomic samples from a range of different sports fields, thus including sports with both high anaerobic and aerobic loads, such as marathon athletes as well as cyclists and rugby players [19, 22]. In addition, 164 metagenomic samples from healthy sedentary adults [23] were included in the study as a control group as well as 69 metagenomic samples of individuals identified as moderate athletes [24, 25].

Selected data led to a total of 418 shotgun metagenomic samples of athletes, sedentary and moderate athletes, supported with categorical (qualitative) physical activity-related metadata derived from their original studies (Table S1). To avoid data analysis biases, such data were re-analyzed following a common bioinformatic pipeline, i.e., METAnnotatorX2 [20]. All the metagenomic datasets showed an average of 3,100,774 reads per sample after Quality and Homo Sapiens filtering steps (Table S1).

Taxonomic features associated with agonistic physical activity.

The first step in the meta-analysis focused on performing descriptive analyses to correlate athletes, moderate athletes, and sedentary category (related to high, average and low physical activity) with the microbial taxonomic profiles, aiming to trace potential key microbial markers related to agonistic sport activity. Processing of all SRA samples through METAnnotatorX2 software (see the "Materials and methods" section for more details) allowed to retrieve of the taxonomic profiles of each analyzed metagenomic dataset with species-level accuracy [26] (Table S2). Furthermore, a hierarchical clustering analysis (HCL) was performed with an ideal number of centroids for the identification of the cluster that was extracted through a Silhouette analysis [27] (Figure S1).

The HCL analyses identified a total of eight taxonomic clusters, named formally Physical activity level Community State Type (PCST) from PCST_1 to PCST_8, each characterized by a unique and recurring average bacterial composition profile (Table S3) (Figure S2). Notably, PCST_3, PCST_7, and PCST_8 represent clusters identified prevalently in the gut microbiomes of athletes and moderate athletes, and their sum represents 77.8%, 100%, and 91.5% of the predicted samples, respectively. In contrast, PCST_1, PCST_4, and PCST_5 were mainly found in the gut microbiomes of sedentary samples (144 out of 166) (Fig. 1a) (Table (Table11).



g)

Spearman Correlation

EC_Cluster	EC	Corr. Value	Asymtoptic P. value	Adj. P. value
EFC_1	PCST_1	0.332567375	2.98E-12	4.99E-12
EFC_1	PCST_5	0.383716943	4.44E-16	9.67E-16
EFC_1	PCST_8	-0.430334948	0.00E+00	0.00E+00
EFC_2	PCST_4	0.587509009	0.00E+00	0.00E+00
EFC_2	PCST_8	-0.257715472	9.10E-08	9.50E-08
EFC_4	PCST_1	-0.280028513	5.71E-09	6.78E-09
EFC_4	PCST_3	0.256344994	1.07E-07	1.10E-07
EFC_4	PCST_4	-0.427676178	0.00E+00	0.00E+00
EFC_4	PCST_5	-0.340622108	8.16E-13	1.43E-12
EFC_4	PCST_8	0.623144997	0.00E+00	0.00E+00

Figure 1

Fig. 1. Samples subdivision between PCST and EFC clusters. In a, the PCST compositions in sample type (athlete, sedentary and moderate athlete) is reported, while in b, the total sample subdivision between the PCSTs is reported. Following the same logic, in c, the EFC compositions in samples type (athlete, sedentary, and moderate athlete) are reported, while in d, the total sample subdivision between the EFCs is reported. e The PCST distribution inside the EFC clusters. f The EFC correlation percentage with EC-Numbers. Finally, in g, the correlation score between EFCs and PCST clusters is reported

Table 1

PCSTs detailed samples subdivision

	PCST_1	PCST_2	PCST_3	PCST_4	PCST_5	PCST_6	PCST_7	PCST_8
Athlete	18	7	28	21	17	1	17	76
Control	31	3	8	49	64	1	0	10
Moderate	11	1	0	15	10	0	1	31
	PCST_1	**PCST_2**	**PCST_3**	**PCST_4**	**PCST_5**	**PCST_6**	**PCST_7**	**PCST_8**
Athlete	30.0%	63.6%	77.8%	24.7%	18.7%	50.0%	94.4%	65.0%
Control	51.7%	27.3%	22.2%	57.6%	70.3%	50.0%	0.0%	8.5%
Moderate	18.3%	9.1%	0.0%	17.6%	11.0%	0.0%	5.6%	26.5%

Notably, PCST_2 and PCST_6 contain less than 15 metagenomic samples, so they were excluded from our analysis because they are outliers, representing uncommon gut microbiota populations with limited statistical relevance (Figure_1, Panel a - b) (Table_1).

Subsequently, we obtained the eigenvalues from the Bray-Curtis dissimilarity matrix, running a Principal Coordinate Analysis (PCoA) to analyze the beta-diversity between the samples (Figure_2).

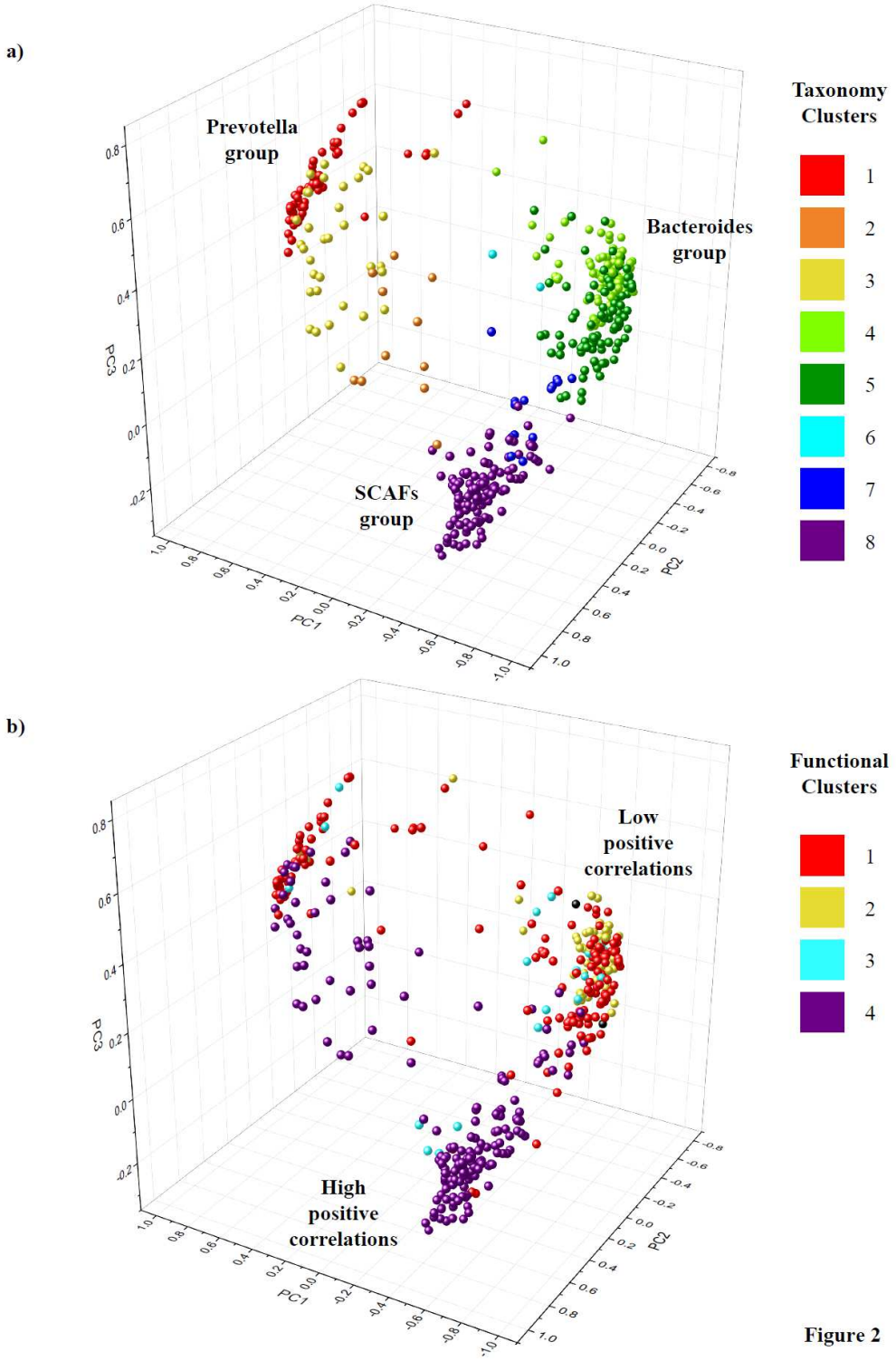

a)
Prevotella group
Bacteroides group
SCAFs group
PC1
PC2
PC3
Taxonomy Clusters
1
2
3
4
5
6
7
8
b)
Low positive correlations
High positive correlations
PC1
PC2
PC3
Functional Clusters
1
2
3
4

Figure 2

Fig. 2. Beta diversity separations of samples based on their compositions and metadata. a The principal coordinate subdivision of the metagenomic samples, based on Bray–Curtis's dissimilarity matrix of taxonomical composition, and subdivided for PCST clusters, with color scheme reported in legend. b The principal coordinate subdivision of samples, based on Bray–Curtis's dissimilarity matrix of taxonomical composition, and subdivided for EFC clusters, with color reported in legend

Through the PCoA analysis, we found that the eight Physical activity Community State Type (PCST) sub-divide samples confirming the marked differences between the taxonomical composition of the different PCSTs (Fig. 2). Furthermore, these data revealed a substantial separation between athletes and sedentary individuals based on their gut microbiota taxonomical composition.

Remarkably, athlete-representative clusters identified based on the distribution of athlete's samples (Fig. 1a) (Table (Table1),1), i.e., PCST_3, PCST_7 and PCST_8, shared a high occurrence of short fatty acid-producing microbial species (SCFAs producers), which distinguish them from the other taxonomic clusters analyzed (Mann–Whitney U adj. P-value < 0.05) (Table S3) (Table S4), thus confirming previous observations [19].

Bacterial SCFAs producers statistically associated to athletes' samples (Mann–Whitney U adj. P-value < 0.05) (Table S4) include *Eubacterium* rectale (3.5 to 11.4% in average relative abundance), *Faecalibacterium prausnitzi* (4.5 to 8.2% in average relative abundance), and other unclassified *Faecalibacterium* species (4.5 to 9.5% in average relative abundance) (Figure S2) (Table S3). Additionally, it has been identified also other microbial species that are potentially involved in the synthesis of SCFAs, i.e., *Ruminococcus bromii* (0.4 to 3.4% in average relative abundance) but also putatively novel unclassified species of *Eubacterium*

(1.3 to 2.4% in average relative abundance) and *Ruminococcus* species (1.5 to 3.4% in average relative abundance) (Mann–Whitney U adj. P-value < 0.05) (Table S4) (Figure S2) (Table S3). Altogether, the above-described bacterial taxa make up the "core" of SCFAs producers relating to athletes. Notably, PCST_3 also showed the presence of another SCFAs bacterial producer in addition to the above-mentioned "core," i.e., *Prevotella*, and more specifically the dominant specie Prevotella copri (21.7%). Nevertheless, *Prevotella* is present also in the Sedentary-related PCST_1 (Kruskal–Wallis adj. P-value < 0.05) (Table S4) (Figure S2) (Table S3). *Prevotella* genus can act as an important microbial producer and consumer of SCFAs, but it has also been associated with various human inflammatory states [28, 29].

Intriguingly, all the PCST clusters containing the most prevalent SCFAs producers related to the genera *Faecalibacterium*, *Eubacterium*, and *Ruminococcus* were primarily identified in the gut microbiomes of athletes, thus reinforcing the previous notion that correlate SCFAs production to physical activity and the diet related to agonistic sports regimes.

Intriguingly, all PCST clusters containing the most prevalent SCFA producers of *Faecalibacterium*, *Eubacterium*, and *Ruminococcus* genera were identified primarily in the gut microbiomes of athletes, thus reinforcing the previous notion that SCFA production is higher in athletes compared to the other individuals (Fig. 1g) (Table S3).

Functional analysis of potential-encoding enzymatic profiles.

While SCFAs production has been extensively investigated for its impact on human health with a range of benefits [30–32], our current scientific understandings of the microbial metabolism leading to the production of secondary compounds involves thousands of enzymatic reactions encompassing catabolic and anabolic pathways, which may be responsible of the athletes' performance and wellbeing. Hence, we performed a functional analysis of the 418 gut microbiomes aimed to identify the enzymatic pathways related to the production of chemical compounds that the scientific literature indicated as able to contribute to the human health by improving physical performances and quality of life. In this framework, METAnnotatorX2 was exploited to retrieve microbially based enzymatic profiles based on the MetaCyc database. Subsequently, a Bray–Curtis distance matrix was generated based on the enzymatic potential of each sample, in order to normalize the results and finally obtain a beta-diversity score (Table S2) (Table S5) that was employed for a hierarchical clustering (HCL) analysis.

We obtained a total of four enzymatic functional clusters (EFC) present in the pool of the analyzed samples, named EFC_1, EFC_2, EFC_3, and EFC_4 (Fig. 1) (Table S3). EFC_1 and EFC_4 represented the most populated clusters, comprising 38.8% and 42.1% of the total pool of samples. On the other hand, clusters EFC_2 and EFC_3 encompassed less frequent enzymatic profiles, including only 14.6 and 4.5% of the metagenomic samples, respectively (Fig. 1) (Figure S2). So, the latter clusters were excluded from further analysis, and we focused only on the most representative functional profiles.

Notably, EFC_4 was composed of 72% of athlete and 18% of moderate athlete, while EFC_1 included 63% of sedentary and 23.7% of moderate athlete (Table (Table2)2) (Fig. 1).

Table 2

EFCs detailed samples subdivision

	EFC_1	EFC_2	EFC_3	EFC_4
Athlete	22	18	18	127
Moderate	38	0	0	31
Control	102	43	1	18
	EFC_1	**EFC_2**	**EFC_3**	**EFC_4**
Athlete	13.58%	29.51%	94.74%	72.16%
Moderate	23.46%	0.00%	0.00%	17.61%
Control	62.96%	70.49%	5.26%	10.23%

Intriguingly, metagenomic samples belonging to moderate athletes were evenly distributed between EFC_1 and EFC_4, highlighting how non-intense or non-prolonged physical activity leads the samples to have in-between enzymatic profiles, an assumption validated by PERMANOVA analysis (adj. P-value < 0.001) (Table S6) (Fig. 1). Therefore, it can be extrapolated how EFC_4 is the most frequent enzymatic profile in the gut microbiome of athletes while EFC_1 is the most common in the gut microbiome of sedentary individuals. Thus, these findings support the strong association between athletes and the gut microbiota composition previously described (Fig. 1a) and highlight another association between athletes and the microbial-based enzymatic profiles (Fig. 1c) (Table (Table22).

Furthermore, we correlated the categorical data deriving from microbial enzymatic clusters (EFCs) with the taxonomic data (PCSTs) to obtain a complete overview of the taxonomic-enzymatic relationships. Such analyses highlighted that EFC_4 correlates with PCST_3, PCST_7, and PCST_8 (Spearman asymptotic adj. P-value < 0.005) (Fig. 1g), i.e., the clusters containing the SCFAs-producing bacteria "core" previously defined (Figure S2). Intriguingly, 60.8% of EFC_4 is composed of metagenomic samples belonging to PCST_8, which is the taxonomical cluster with the highest presence of Eubacterium rectale as well as Faecalibacterium prausnitzii and other Faecalibacterium spp. (Figure S2) (Fig. 1).

Instead, the EFC_1 cluster correlates with PCST_1 and PCST_5 clusters (Spearman asymptotic adj. P-value < 0.005), mainly dominated by the genera Prevotella, Bacteroidetes, and Alistipes, with species such as Prevotella copri, Bacteroides uniformis, Bacteroides stercoris, and Alistipes uniformis (Figure S2) (Fig. 1g).

In addition, we further detailed each EFC-cluster's association with each enzymatic reaction profiled, following the Enzyme Commission nomenclature (EC-Numbers) [33]. For this purpose, only those ECs displaying a prevalence > 10% were considered, for a total of 1604 EC numbers (Table S3) (Table S7). Unexpectedly, EFC_4 displays 725 of positive correlations (80% of its total statistically significative correlation, Spearman asymptotic adj. P-value < 0.05) with the retained ECs, showing a large gap compared to the EFC_1, which on average showed a total of only 79 positive correlations (25% of its total statistically significative correlation) (Table S7) (Fig. 1f). These findings, clearly corroborate what preliminary observed in a previous study [34] encompassing a

small cohort of individuals analyzed with a less accurate metagenomic approach such as the 16S rRNA gene microbial profiling. Remarkably, the shotgun metagenomic approach allowed us also to explore in detail the metabolic relevance of enzymatic reactions positively correlated with physical activity.

Characterization of microbial biosynthetic metabolisms associated with physical activity.

The two main enzymatic clusters, i.e., EFC_1 and EFC_4, were also exploited to investigate those enzymes involved in the anabolism of key metabolites known to impact on host's health by the recent scientific literature [35, 36]. In this context, a selection of EC numbers was manually investigated for their possible relevance and were named high biological impact synthases (HBIS) (Table S8). Notably, these ECs were selected based on information reported in the MetaCyc database and cited literature data [35, 36] (Table S8).

A comparison of the enzymatic profiles of HBIS between the two groups revealed 66 HBIS positively correlated with cluster EFC_4 (representing the most common enzymatic profile of athletes) and only 10 with cluster EFC_1 (representing the most common functional profile of sedentary individuals) (Table S8). Therefore, the EFC_1 enzyme cluster displays a lower HBIS production potential than EFC_4, highlighting how the microbiota of athletes can potentially encode for a much wider range of microbial metabolites with an important impact on health and physical performance.

In detail, between the 14 HBIS positively related to EFC_1 there are EC related mainly to vitamin biosynthesis, but also related to flavodoxin precursor, a well-

known phosphoantigen also required by many pathogens to survive [37, 38] (Table (Table33).

Table 3

HBIS positively correlated to EFC_1 and manually identified with MetaCyc database

Spearman correlation (adj. *P*-value < 0.05)					
EC name	**EC number**	**EFC_4**	**EFC_1**	**Related product effects**	**Ref**
Isochorismate synthase	5.4.4.2	− 0.541970	0.311153	Production of the precursor of vitamin K_2	[39]
1,4-dihydroxy-2-naphthoyl-CoA synthase	4.1.3.36	− 0.543416	0.316668	Production of the precursor of vitamin K_2	[40]
Quinolinate synthase	2.5.1.72	-	0.266977	Precursor of niacin and indirectly of vitamin B_3	[41]
(E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase	1.17.7.3	− 0.598568	0.333534	Production of phosphoantigen	[37]

In contrast, among the 73 positive correlations between EFC_4 and HBIS, we extracted and focused on eight enzymes related to the enhancement of sports performance and the increase of life span through the reduction of the onset of cardiovascular diseases and tumors. Among the enzymes selected, there is also an enzyme involved in the production of the heme group and therefore in the regeneration and production of new blood cells (Table S8) (Table 4).

Table 4

HBIS positively correlated to EFC_4 and manually identified with MetaCyc database

Spearman correlation (adj. *P*-value < 0.05)					
EC name	**EC number**	**EFC_4**	**EFC_1**	**Related products effects**	**Ref**
Spermidine synthase	2.5.1.16	0.764993	− 0.340411	Reduction in mortality due to cardiovascular disease	[42, 43]
Porphobilinogen synthase	4.2.1.24	0.590898	-	Heme biosynthesis	[44]
Mycothiol synthase	2.3.1.189	0.771842	− 0.319983	Antibacterial and antitumoral properties	[45, 46]
Hydrogenobyrinic acid a,c-diamide synthase	6.3.5.9	0.642982	− 0.284009	Coenzyme B12 (cobalamin) biosynthesis	[47, 48]
Cystathionine gamma-synthase	2.5.1.48	0.712590	− 0.256458	Energy metabolism, muscle performance, antioxidant	[49–54]
Glutamate synthase (NADPH)	1.4.1.13	0.623222	-	Excitatory neurotransmitter, homocysteine balance	[55–58]
Asparagine synthase	6.3.5.4	0.645870	-	Excitatory neurotransmitter, homocysteine balance	[55–58]
Glutaminyl-tRNA synthase	6.3.5.7	0.768508	− 0.284009	Excitatory neurotransmitter, homocysteine balance	[55–58]

Additionally, EC numbers related to the production of sulfur amino acids and molecules like glutathione (GSH) and taurine were correlated positively with EFC_4, potentially enhancing the reduction of oxidative-cellular damage and boosting muscular performance (Table S8) (Table 4).

Altogether, these results evidenced that the microbiomes of the samples belonging to athlete's category are characterized by a higher abundance of biosynthetic enzymes involved in the production of a wide range of compounds (Fig. 3).

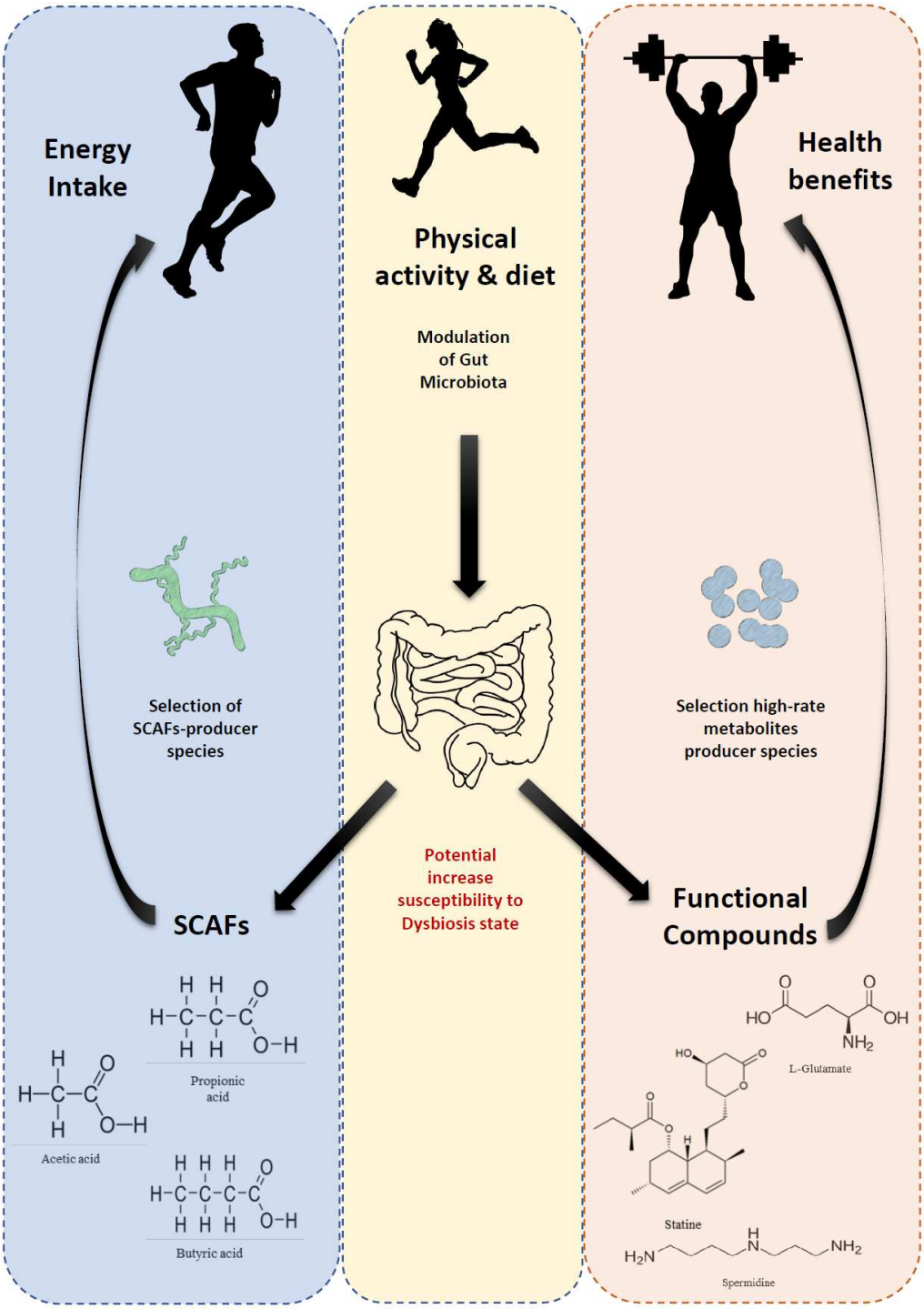

Energy Intake
Physical activity & diet
Health benefits
Modulation of Gut Microbiota
Selection of SCAFs-producer species
Selection high-rate metabolites producer species
Potential increase susceptibility to Dysbiosis state
SCAFs
Functional Compounds
Propionic acid
Acetic acid
Butyric acid
L-Glutamate
Statine
Spermidine

Figure_3

Fig. 3. Schematic representation of the project aims and key points. The modulation effect that physical activity can exert on gut microbiota and vice versa the effect that gut microbiota can exert on human health and performance. Some of the main compounds produced by SCFAs producers are reported with name and structural formula.

Associations between HBIS and Core microbial taxa.

We performed a taxonomic EC back-tracking analysis to investigate further the main bacterial taxa responsible for the above-reported enzymatic reactions associated with physical activity. This approach aims to identify the bacterial species that can potentially produce the nine HBIS positively correlated to the EFC_4 above-discussed.

As expected, the "core" of SCFAs producers found in athlete metagenomic samples, such as *Faecalibacterium prausnitzii*, *Eubacterium rectale*, and *Blautia wexlerae* and a set of minor representative species of *Faecalibacterium*, *Eubacterium*, *Ruminococcus*, and *Blautia* genera act as major microbial producers of the nine enzymatic reactions previously highlighted as possessing a high putative health interest in the EFC_4 cluster (Table S9). In detail, six EC classes (EC 6.3.5.9, 6.3.5.7, 4.2.1.24, 2.5.1.16, 6.3.5.4, and 2.3.1.189) resulted to be produced primarily by the above-identified "core" of SCFAs producers. Moreover, EC 1.4.1.13, a glutamate synthase (NADPH), was found to be produced more specifically by *Faecalibacterium prausnitzii*, *Eubacterium rectale*, and other *Faecalibacterium* species (Table S9). In contrast, EC 2.5.1.48, which encompasses a cystathionine gamma-synthase, was predicted to be produced by a more variegated number of species, including *Anaerostipes*, *Ruminooccus*, and *Coprococcus* species, along with *Bifidobacterium adolescentis* (Table S9).

Intriguingly, these data revealed clear associations between specific functional features and microbial taxa harbored by the intestinal environment of athletes.

Conclusion

With the purpose of analyzing the intricate relationships between the gut microbiome and athletes' related lifestyle (multifactorial metadata including training, diet, and stress), we statistically analyzed 418 metagenomic samples divided into athlete, sedentary, and moderate athletes. As a result of taxonomical profiling, we identified a correlation between gut microbial profiles and athlete's category, as evidenced by a recurrent microbial pattern defined primarily by SCFAs microbial producers including *Faecalibacterium*, *Eubacterium*, *Blautia*, and *Ruminococcus* species, which are statistically associated to athletes' samples (Table S4).

In addition, subsequent functional analysis showed the presence of two major enzymatic functional clusters (EFCs), one strongly associated with the presence of sedentary individuals and one with athletes, thus corroborating the differences previously seen at species-taxonomical level between the two types of samples (athletic and sedentary subjects). Intriguing, the EFC related to athletes was positively linked to 752 enzymes (EC numbers) and 73 high biological impact synthases (HIBS), a subset of manually identified biosynthetic reactions. In contrast, the EFC related to sedentary resulted in being positively linked only to 105 EC numbers and 14 HBIS, highlighting the reduced ability of sedentary' gut microbiota to affect the host health through the production of secondary metabolites. Furthermore, the correlation of the enzymatic potential with species-

level microbial profiles evidenced how additional microbial taxa may be implicated in the biosynthesis of compounds of high biological interest.
Remarkably, these data highlighted how the athletes' related lifestyle represent a multifactorial ecological pressure that modulate the gut microbiota, reshaping it in favor of bacterial species with a higher enzymatic potential impacting the host's health and muscular performances. Additionally, all these results pointed out how the bacterial species commonly considered core SCFAs producers are also implicated in the production of a much wider and variegated range of potentially high functional impact molecules, which will require a precise characterization in future population studies.

Materials and methods

Metagenomic sample collection.

A set of 418 shotgun metagenomic data were retrieved from the National Center of Biotechnology Information (NCBI) Sequence Read Archive (SRA) database. The terms used to inspect the scientific literature include athlete, gut microbiota, IBD, SCFA, sedentary, performance, physical activity. For the selection of the optimal Bioprojects for this study, we used various criteria, such as the selection of healthy samples, the sequencing technology, the minimum number of reads available and finally the completeness of the metadata regarding athlete and sedentary categories. All Bioprojects have been manually checked to ensure that minimum criteria were met. In detail, each metagenomic dataset possess a minimum of 10000 reads, according to the minimum sequencing depth required to METAnnotatorX2 for obtain high quality taxonomical profiles [26].

Accordingly, we collected shotgun metagenomics sequences and associated metadata from six different studies (PRJEB15388, PRJEB28338, PRJEB32794, PRJNA472785, PRJNA305507, PRJEB20054). The selection of six different sources (Bioprojects) of raw data sequenced through illumina technology allowed reduced selection bias. Additionally, this selection was performed to obtain a comparable number of samples between athletes and controls. In detail, 185 samples corresponded to athlete gut microbiomes, 69 to moderate athlete and 164 were from healthy sedentary individuals (Table_S1). The athletes and the sedentary categories were defined by metadata originating from their original scientific articles and Bioprojects. Moderate athletes instead refer to athletes who have performed competitive activity only for a short time window (high school athletes) or without reaching the higher categories [therefore CAT 1 (semi-professional) vs. PRO athletes]. Specifically, between the 69 moderate athletes' samples were included time-longitudinal samples belonging to bioprojects PRJNA472785 and PRJNA305507 to increase the robustness of the analysis regarding the group composed by moderate athletes. Thus, the small group of moderate athletes was used to compare and validate the distribution of the two main analysis groups (Athletes and Controls). Additional metadata regarding physical status, type of sport performed and other miscellaneous are reported along with the SRA name in Table S1. All available metadata regarding the metagenomic samples (mainly athletic and sedentary designation) were retrieved from the bioprojects related to the samples.

Metagenomics data processing, taxonomic profiling and functional analysis. Each metagenomic datasets were filtered to remove reads with a base sequence quality of < 25 (score obtained from FastQC software for Illumina sequencing), and to retain reads with a length of > 149 bp. Taxonomic and functional profiling of reads resulting from quality and *Homo sapiens* filtering was performed with the METAnnotatorX2 bioinformatics platform [26,59]. Within the METAnnotatorX2 pipeline, MegaBLAST [60] was employed for taxonomic classification of each metagenomic read, using a curated non-redundant sequence database of genomes Each metagenomic datasets were filtered to remove reads with a base sequence quality of < 25 (score obtained from FastQC software for Illumina sequencing) and to retain reads with a length of > 149 bp. Taxonomic and functional profiling of reads resulting from quality and Homo sapiens filtering was performed with the METAnnotatorX2 bioinformatics platform [26, 59]. Within the METAnnotatorX2 pipeline, MegaBLAST [60] was employed for taxonomic classification of each metagenomic read, using a curated non-redundant sequence database of genomes retrieved from NCBI servers and manually selected. The generation of the taxonomical database was reported in detail by Milani et al. [26] and periodically updated (every 6 months). Reads with a nucleotide identity of > 94% to reference genomes are classified at the species level, while reads with a lower percentage identity are classified at the genus level as undefined species. The functional enzymatic classification of each metagenomic read was performed through DIAMOND [61], employing a curated non-redundant sequence database of EC number sequence created employing the MetaCyc database [62]. DIAMOND parameters used for this analysis were as default chosen by the METAnnotatorX2 pipeline using up to 5,000,000 reads (–

query-cover 80, –evalue 0.00000001, and –max-target-seqs 1). Taxonomic EC back-tracking analysis was performed using METAnnotatorX2 -x ec_taxonomy function. This function allowed to retrieve the bacterial species related to the production of a selected list of enzymatic codes.

For the analyses that required the use of R software, version R-4.1.2 was used, along the version RStudio-2021.09.2–382 of R Studios and rtools40v2-x86_64 of rtools.

Similarities between samples (beta-diversity) were calculated using the Bray–Curtis distance matrix based on species relative abundance, using the vegdist function (from vegan_2.5–7) on R-Studios (RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL.). The range of similarities is calculated between values 0 and 1. PCoA representation of beta-diversity was performed using ORIGIN 2021b (https://www.originlab.com/2021).

In the PCoA, each dot represented a sample, distributed in tridimensional space according to its bacterial composition, i.e., eigenvalues scores. The hierarchical clustering analysis (HCA) of samples, performed on ORIGIN 2021b, was achieved employing Bray–Curtis matrix using Pearson correlation as a distance metric and the sum square of distances and furthest neighbor for clustering methods. The optimal number of clusters was defined through a Silhouette analysis [27] performed on ORIGIN 2021b. The data obtained was represented by a vertical dendrogram.

Statistical analysis.

ORIGIN 2021b (https://www.originlab.com/2021), IBM SPSS statistics software (version 25) (www.ibm.com/software/it/analytics/spss/) and R-Studios were used to compute statistical analyses. PERMANOVA analyses were performed on R-studios using 999 permutations to assess p-values for population differences in PCoA analyses. In detail, input data was preprocessed and transformed in a Bray Curtis dissimilarity matrix with vegdist function (from vegan_2.5-7) and the PERMANOVA analysis was performed with adonis2 package (from vegan_2.5-7). Non-parametric Kruskal-Wallis's test was performed on SPSS software using PCSTs subdivision as group criteria. In addition, a Pairwise Post-hoc Analysis was performed for the Kruskal-Wallis's analysis, using Bonferroni correction for the FDR adj. p value. Non-parametric Mann-Whitney U Test was performed on SPSS software using PCST_1, PCST_4 and PCST_5 as Group 1 and PCA_3, PCST_7 and PCST_8 as Group 2. Spearman correlation was performed with rcorr function (from Hmisc_4.6-0), and only statistical significative results with correlation score greater than 0.25 or minor of -0.25 were retained. The eigenvalues were retrieved from the Bray Curtis dissimilarity matrix with the use of prcomp function (from base package stats) and the get_pca function (from factoextra_1.0.7). All the raw p-value with the exclusion of Kruskal-Wallis's Pairwise Post-hoc were subjected to FDR correction using Benjamini-Hochberg [63] approach on R-studios through p.adjust function (from base package stats).

Ethics approval and Consent to participate.

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data can be retrieved from NCBI SRA repository through their SRA Accession Number reported in Table_S1.

Competing interests

The authors declare that they have no competing interests that might be perceived to influence the results and/or discussion reported in this paper.

Funding

Not applicable.

Authors' contributions

F.F. performed bioinformatics analyses and wrote the manuscript; C.M. validated the bioinformatics analyses and edited the manuscript; L.M., G.A.L., C.T., G.L, and G.A. managed the metadata and data results; F.T. supervised the project and edited the manuscript; M.V. supervised the project and designed the study.

Acknowledgments

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

References

1. Cella V, Bimonte VM, Sabato C, Paoli A, Baldari C, Campanella M, et al. Nutrition and physical activity-induced changes in gut microbiota: possible implications for human health and athletic performance. Foods. 2021;10(12)3075. 10.3390/foods10123075.
2. de Vos WM, Tilg H, Van Hul M, Cani PD. Gut microbiome and health: mechanistic insights. Gut Gut. 2022;71:1020–1032. doi: 10.1136/gutjnl-2021-326789.
3. Vernocchi P, Chierico F Del, Putignani L. Gut microbiota metabolism and interaction with food components. Int J Mol Sci. 2020;21(10)3688. 10.3390/ijms21103688.
4. Marchesi JR, Adams DH, Fava F, Hermes GDA, Hirschfield GM, Hold G, et al. The gut microbiota and host health: a new clinical frontier. Gut BMJ Publishing Group. 2016;65:330–339.
5. Kiani AK, Bonetti G, Donato K, Bertelli M. Dietary supplements for intestinal inflammation. J Prev Med Hyg. 2022;63(2 Suppl 3):E214–20. 10.15167/2421-4248/jpmh2022.63.2S3.2763.
6. Chicco F, Magrì S, Cingolani A, Paduano D, Pesenti M, Zara F, et al. Multidimensional impact of Mediterranean diet on IBD patients. Inflamm Bowel Dis. Inflamm Bowel Dis; 2021 [cited 15 Jan 2023];27:1–9. Available from: https://pubmed.ncbi.nlm.nih.gov/32440680/
7. Raoul P, Cintoni M, Palombaro M, Basso L, Rinninella E, Gasbarrini A, et al. Food Additives, a key environmental factor in the development of IBD through gut dysbiosis. Microorganisms. 2022;10(1):167. 10.3390/microorganisms10010167.
8. Rinninella E, Cintoni M, Raoul P, Lopetuso LR, Scaldaferri F, Pulcini G, et al. Food components and dietary habits: keys for a healthy gut microbiota composition. Nutrients. 2019;11(10):2393. doi: 10.3390/nu11102393.
9. Clark A, Mach N. Exercise-induced stress behavior, gut-microbiota-brain axis and diet: a systematic review for athletes. J Int Soc Sports Nutr [Internet]. J Int Soc Sports Nutr; 2016 [cited 14 Jan 2023];13. Available from: https://pubmed.ncbi.nlm.nih.gov/27924137/
10. Sommer F, Anderson JM, Bharti R, Raes J, Rosenstiel P. The resilience of the intestinal microbiota influences health and disease. Nat Rev Microbiol Nat Rev Microbiol. 2017;15:630–638. doi: 10.1038/nrmicro.2017.58.

11. Clark A, Mach N. Exercise-induced stress behavior, gut-microbiota-brain axis and diet: a systematic review for athletes. J Int Soc Sports Nutr. 2016;13:43. doi: 10.1186/s12970-016-0155-6.
12. Suryani D, Subhan Alfaqih M, Gunadi JW, Sylviana N, Goenawan H, Megantara I, et al. Type, intensity, and duration of exercise as regulator of gut microbiome profile. Curr Sports Med Rep. 2022;21:84–91. doi: 10.1249/JSR.0000000000000940.
13. Morishima S, Aoi W, Kawamura A, Kawase T, Takagi T, Naito Y, et al. Intensive, prolonged exercise seemingly causes gut dysbiosis in female endurance runners. J Clin Biochem Nutr. 2021;68:253. doi: 10.3164/jcbn.20-131.
14. Morishima S, Oda N, Ikeda H, Segawa T, Oda M, Tsukahara T, et al. Altered fecal microbiotas and organic acid concentrations indicate possible gut dysbiosis in university rugby players: An observational study. Microorganisms. 2021;9(8)1687. 10.3390/microorganisms9081687.
15. Bonomini-Gnutzmann R, Plaza-Díaz J, Jorquera-Aguilera C, Rodríguez-Rodríguez A, Rodríguez-Rodríguez F. Effect of intensity and duration of exercise on gut microbiota in humans: a systematic review. Int J Environ Res Public Health. 2022;19(15):9518. doi: 10.3390/ijerph19159518.
16. Moreno-Pérez D, Bressa C, Bailén M, Hamed-Bousdar S, Naclerio F, Carmona M, et al. Effect of a protein supplement on the gut microbiota of endurance athletes: a randomized, controlled, doublE-BLIND PILOT STUDY. Nutrients. 2018;10(3):337. doi: 10.3390/nu10030337.
17. Imdad S, Lim W, Kim J-H, Kang C. Intertwined relationship of mitochondrial metabolism, gut microbiome and exercise potential. Int J Mol Sci. 2022;23:2679. doi: 10.3390/ijms23052679.
18. Hughes RL, Holscher HD. Fueling gut microbes: a review of the interaction between diet, exercise, and the gut microbiota in athletes. Adv Nutr. 2021;12:2190. doi: 10.1093/advances/nmab077.
19. Barton W, Penney NC, Cronin O, Garcia-Perez I, Molloy MG, Holmes E, et al. The microbiome of professional athletes differs from that of more sedentary subjects in composition and particularly at the functional metabolic level. Gut [Internet]. Gut; 2018 [cited 30 Jun 2022];67:625–33. Available from: https://pubmed.ncbi.nlm.nih.gov/28360096/
20. Markowiak-Kopeć P, Śliżewska K. The effect of probiotics on the production of short-chain

fatty acids by human intestinal microbiome. Nutrients. 2020;12(4)1107. 10.3390/nu12041107.

21. Feng W, Liu J, Cheng H, Zhang D, Tan Y, Peng C. Dietary compounds in modulation of gut microbiota-derived metabolites. Front Nutr. 2022;9:1564. doi: 10.3389/fnut.2022.939571.
22. O'Donovan CM, Madigan SM, Garcia-Perez I, Rankin A, O' Sullivan O, Cotter PD. Distinct microbiome composition and metabolome exists across subgroups of elite Irish athletes. J Sci Med Sport. 2020 [cited 6 Apr 2022];23:63–8. Available from: https://pubmed.ncbi.nlm.nih.gov/31558359/
23. Cronin O, Barton W, Skuse P, Penney NC, Garcia-Perez I, Murphy EF, et al. A prospective metagenomic and metabolomic analysis of the impact of exercise and/or whey protein supplementation on the gut microbiome of sedentary adults. mSystems. 2018;3(3)e00044–18. 10.1128/mSystems.00044-18.
24. Petersen LM, Bautista EJ, Nguyen H, Hanson BM, Chen L, Lek SH, et al. Community characteristics of the gut microbiomes of competitive cyclists. Microbiome. 2017;5(1)98. 10.1186/s40168-017-0320-4.
25. Scheiman J, Luber JM, Chavkin TA, MacDonald T, Tung A, Pham LD, et al. Meta'omic analysis of elite athletes identifies a performance-enhancing microbe that functions via lactate metabolism. Nat Med. 2019;25:1104. doi: 10.1038/s41591-019-0485-4.
26. 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. Arumugam M, editor. mSystems. 2021;6(3):e0058321. 10.1128/mSystems.00583-21.
27. Lengyel A, Botta-Dukát Z. Silhouette width using generalized mean-a flexible method for assessing clustering efficiency. Ecol Evol Ecol Evol. 2019;9:13231–13243. doi: 10.1002/ece3.5774.
28. Larsen JM. The immune response to Prevotella bacteria in chronic inflammatory disease. Immunology Immunology. 2017;151:363–374.
29. Chen C, Fang S, Wei H, He M, Fu H, Xiong X, et al. Prevotella copri increases fat accumulation in pigs fed with formula diets. Microbiome. 2021;9(1)175. 10.1186/s40168-021-01110-0.
30. Luu M, Monning H, Visekruna A. Exploring the molecular mechanisms underlying the

protective effects of microbial SCFAs on intestinal tolerance and food allergy. Front Immunol. 2020;11:1225. doi: 10.3389/fimmu.2020.01225.

31. Dalile B, Van Oudenhove L, Vervliet B, Verbeke K. The role of short-chain fatty acids in microbiota-gut-brain communication. Nat Rev Gastroenterol Hepatol. 2019;16:461–78. doi: 10.1038/s41575-019-0157-3.
32. Ticinesi A, Mancabelli L, Tagliaferri S, Nouvenne A, Milani C, Del Rio D, et al. The gut-muscle axis in older subjects with low muscle mass and performance: a proof of concept study exploring fecal microbiota composition and function with shotgun metagenomics sequencing. Int J Mol Sci. 2020;21:1–16. doi: 10.3390/ijms21238946.
33. Enzyme Nomenclature. [cited 2022 Apr 6]. Available from: https://iubmb.qmul.ac.uk/enzyme/
34. Barton W, Penney NC, Cronin O, Garcia-Perez I, Molloy MG, Holmes E, et al. The microbiome of professional athletes differs from that of more sedentary subjects in composition and particularly at the functional metabolic level. Gut Gut. 2018;67:625–633. [Google Scholar]
35. Fan Y, Pedersen O. Gut microbiota in human metabolic health and disease. Nature Reviews Microbiology 2020 19:1. Nature Publishing Group; 2020;19:55–71.
36. Shen G, Wu J, Ye BC, Qi N. Gut microbiota-derived metabolites in the development of diseases. Can J Infect Dis Med Microbiol. 2021;2021:6658674. 10.1155/2021/6658674.
37. Sancho J. Flavodoxins: sequence, folding, binding, function and beyond. Cell Mol Life Sci. Cell Mol Life Sci; 2006 [cited 7 Jul 2022];63:855–64. Available from: https://pubmed.ncbi.nlm.nih.gov/16465441/
38. Salillas S, Sancho J. Flavodoxins as novel therapeutic targets against helicobacter pylori and other gastric pathogens. Int J Mol Sci [Internet]. Multidisciplinary Digital Publishing Institute (MDPI); 2020 [cited 14 Jan 2023];21. Available from: /pmc/articles/PMC7084853/
39. Daruwala R, Bhattacharyya DK, Kwon O, Meganathan R. Menaquinone (vitamin K2) biosynthesis: overexpression, purification, and characterization of a new isochorismate synthase from Escherichia coli. J Bacteriol. 1997;179(10):3133–8. doi: 10.1128/jb.179.10.3133-3138.1997.
40. Sun Y, Song H, Li J, Jiang M, Li Y, Zhou J, et al. Active site binding and catalytic role of bicarbonate in 1,4-dihydroxy-2-naphthoyl coenzyme A synthases from vitamin K biosynthetic pathways. Biochemistry. 2012;51:4580–4589. doi: 10.1021/bi300486j.

41. Suo J, Gao Y, Zhang H, Wang G, Cheng H, Hu Y, et al. New insights into the accumulation of vitamin B 3 in Torreya grandis nuts via ethylene induced key gene expression. Food Chem. 2022;371:131050. 10.1016/j.foodchem.2021.131050.

42. Madeo F, Carmona-Gutierrez D, Kepp O, Kroemer G. Spermidine delays aging in humans. Aging (Albany NY) 2018;10:2209. doi: 10.18632/aging.101517.

43. Kiechl S, Pechlaner R, Willeit P, Notdurfter M, Paulweber B, Willeit K, et al. Higher spermidine intake is linked to lower mortality: a prospective population-based study. Am J Clin Nutr; 2018 [cited 14 Jan 2023];108:371–80. Available from: https://pubmed.ncbi.nlm.nih.gov/29955838/

44. Jaffe EK. Porphobilinogen synthase: an equilibrium of different assemblies in human health. Prog Mol Biol Transl Sci. 2020;169:85–104. doi: 10.1016/bs.pmbts.2019.11.003.

45. Lü J, He Q, Huang L, Cai X, Guo W, He J, et al. Accumulation of a bioactive benzoisochromanequinone compound kalafungin by a wild type antitumor-medermycin-producing streptomycete strain. PLoS One. 2015;10:e0117690. doi: 10.1371/journal.pone.0117690.

46. Deng MR, Li Y, Luo X, Zheng XL, Chen Y, Zhang YL, et al. Discovery of mycothiogranaticins from streptomyces vietnamensis GIMV4.0001 and the regulatory effect of mycothiol on the granaticin biosynthesis. Front Chem. 2021;9:802279. 10.3389/fchem.2021.802279.

47. Ryan-Harshman M, Aldoori W. Vitamin B12 and health. Can Fam Phys. 2008;54:536.

48. Boachie J, Adaikalakoteswari A, Gazquez A, Zammit V, Larque E, Saravanan P. Vitamin B12 induces hepatic fatty infiltration through altered fatty acid metabolism. Cell Physiol Biochem. 2021;55:241–255. doi: 10.33594/000000368.

49. Stipanuk MH, Ueki I. Dealing with methionine/homocysteine sulfur: cysteine metabolism to taurine and inorganic sulfur. J Inherit Metab Dis. 2011;34:17. doi: 10.1007/s10545-009-9006-9.

50. Brosnan JT, Brosnan ME. The sulfur-containing amino acids: an overview. J Nutr. 2006;136(6 Suppl):1636S–40S. 10.1093/jn/136.6.1636S.

51. Sbodio JI, Snyder SH, Paul BD. Regulators of the transsulfuration pathway. Br J Pharmacol. 2019;176:583. doi: 10.1111/bph.14446.

52. Wen C, Li F, Zhang L, Duan Y, Guo Q, Wang W, et al. Taurine is involved in energy

metabolism in muscles, adipose tissue, and the liver. Mol Nutr Food Res. 2019;63(2):e1800536. 10.1002/mnfr.201800536.

53. Homma T, Fujii J. Application of glutathione as anti-oxidative and anti-aging drugs. Curr Drug. 2015;16:560–571. doi: 10.2174/1389200216666151015114515.

54. Baliou S, Adamaki M, Ioannou P, Pappa A, Panayiotidis MI, Spandidos DA, et al. Protective role of taurine against oxidative stress (Review) Mol Med Rep. 2021;24(2):605. doi: 10.3892/mmr.2021.12242.

55. Brosnan JT, Brosnan ME. Glutamate: a truly functional amino acid. Amino Acids Amino Acids. 2013;45:413–418. doi: 10.1007/s00726-012-1280-4.

56. Stover PJ, Field MS. Trafficking of intracellular folates. Adv Nutr. 2011;2:325. doi: 10.3945/an.111.000596.

57. Shams A. Folates: an introduction. B-complex vitamins - sources, intakes and novel applications. IntechOpen; 2022.

58. Petroff OAC. GABA and glutamate in the human brain. Neuroscientist. 2002;8(6):562–73. doi: 10.1177/1073858402238515.

59. 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;6(1)145. 10.1186/s40168-018-0527-z.

60. Chen Y, Ye W, Zhang Y, Xu Y. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res. 2015;43:7762–7768. doi: 10.1093/nar/gkv784.

61. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60. doi: 10.1038/nmeth.3176.

62. Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, et al. The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Res [Internet]. Nucleic Acids Res; 2018 [cited 14 Jan 2023];46:D633–9. Available from: https://pubmed.ncbi.nlm.nih.gov/29059334/

63. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological). John Wiley & Sons, Ltd; 1995;57:289–300.

Chapter 5

Investigating the infant gut microbiota in developing countries: worldwide metagenomic meta-analysis involving infants living in sub-urban areas of Côte d'Ivoire.

Federico Fontana, Leonardo Mancabelli, Gabriele Andrea Lugli, Chiara Taracchini, Giulia Alessandri, Giulia Longhi, Rosaria Anzalone, Alice Viappiani, Roch Famo, Marc Brognan, Kouamé Hervé Micondo, Francesca Turroni, Marco Ventura, Rossella D'Alfonso, Christian Milani

The results of this chapter were published in Environmental Microbiology, 2021 Jun 21; https://doi.org/10.1111/1758-2229.12960.

Abstract

In recent decades, infants' gut microbiota has aroused constant scientific interest, primarily due to early- and long-term repercussions on the host health. In this context, nutritional challenges such as those found in less developed countries can influence infants' gut microbiota development, thus generating potentially critical health outcomes. However, comprehensive investigations regarding species-level differences in the infant gut microbiota's composition between urbanized and rural countries are still missing. In this study, 16S rRNA and Shallow Shotgun metagenomics sequencing were exploited to dissect the microbial community's species-level composition of 11 faecal samples collected from infants living in a semi-urban area of Sub-Saharan Africa, i.e. Côte d'Ivoire. Moreover, the generated data were coupled with those retrieved from public available metagenomic repositories, including two rural communities and 13 urban communities of industrialized countries. The meta-analysis led to the identification of Infant Species Community States Type (ISCSTs) and microbial species covariances, which were exploited to reveal key signatures of infants living in rural and semi-urban societies. Remarkably, analysis of rural and semi-urban datasets revealed shifts from ISCSTs prevalent in urbanized populations with putative health implications. Thus, indicating the need for population-wide investigations aimed to define the factors determining such potentially harmful gut microbial communities' signatures.

Introduction

The gastrointestinal tract is commonly colonized by a large variety of microorganisms, which coexist with the host and constitute a complex ecosystem shaped by external stimuli, such as nutrition and drug intake. Human gut microbiota is known to influence the host's health status in different ways (Adak and Khan, 2019). In such a context, it can exert positive influences, but this requires that the microbiota is under a homeostatic equilibrium, or it can negatively affect human health in perturbation of balance between bacterial species, known as dysbiosis that potentially favours pathological conditions (Nagpal and Yamashiro, 2018; Turroni et al., 2020). Furthermore, early dysbiosis in infants can alter the development of the adult gut microbiota, leading to unpleasant mid and long-term effects (Savino et al., 2004; Schirbel and Fiocchi, 2011; Marchesi et al., 2016).

Therefore, it is essential to profile and monitor the bacterial communities to identify significant shifts predisposing to dysbiotic conditions. It is crucial to extend our investigations regarding the infant gut microbiota composition within the first year of life, which is considered a critical window of time to understand the factors protecting later health (Milani et al. 2017a; Cukrowska et al. 2020a). The first major factor impacting the early stages of infant gut microbiota development is the delivery method, i.e. Natural and C-Section (Rutayisire et al., 2016). Feeding with breast or formula milk modulates the intestinal microbial composition of the newborn (De Leoz et al., 2015; Le Doare et al., 2018; Lugli et al., 2020). During weaning, which generally occurs around 6 months after birth, the intestinal microbiota's development is driven by the bacterial colonization from the mother, the environment and the diet (Matamoros et al.,

2013; Gentile and Weir, 2018). The crucial step for intestinal bacterial development is determined by changes in diet. Indeed, novel micro- and macronutrients are available and new substrates start the selection of new bacterial strains based on their metabolic ability (Yadav et al., 2018).

Recently, 12 highly recurrent gut microbial profile structures, known as Infants Community State types (ICSTs), have been identified with genus-level accuracy within the first year of life, of which five associated with lactation and seven with post-weaning gut microbiota development (Mancabelli et al., 2020). However, the ICSTs identified in the available scientific literature are based on samples collected from infants living in urbanized and sub-urbanized countries. Therefore, rural or semi-urban countries have not been assayed in this analysis. Moreover, the latter studies relied on 16S rRNA gene microbial profiling data, resulting in genus-level accuracy.

In this study, a comprehensive meta-analysis of 1109 shotgun metagenomics datasets of infant's faecal samples collected and sequenced in this study or previously released in the framework of published scientific literature was selected, representing a range of geographical regions including 13 urbanized area, two rural areas from Malawi and sub-Saharan Africa as well as one sub-urban area from Côte d'Ivoire collected and submitted to sequencing in this study. The use of shotgun metagenomics data resulted in the identification of novel Community State Types at the species level resolution (Laudadio et al., 2018). These new CSTs, defined in this study as Infants Species Community State Types (ISCSTs), along with species-level profiling covariance analyses, were exploited to explain how the faecal microbial profiles of the rural and semi-urban

communities can be related to the most common ISCSTs observed in urbanized communities.

Results and Discussion

16S rRNA genera profiling analysis of 11 infants from Côte d'Ivoire.

In the framework of this study, 11 faecal samples of infants from Côte d'Ivoire with no reported health problems were collected between 10 and 75 days of life. Notably, these represent the first investigation of the gut microbiota community harboured by infants living in this developing country. Nine out of 11 of these infants were fed with a diet based on mother's milk, while only two were fed with a mixed formula (Table S1). In addition, 1098 faecal samples, corresponding to urbanized and rural infants aged between 1 day and 1 year, were selected from 16 different publicly available shotgun metagenomics datasets obtained by Illumina sequencing to cover both pre- and post-weaning gut microbiota development (Table S1; Fig. S1).

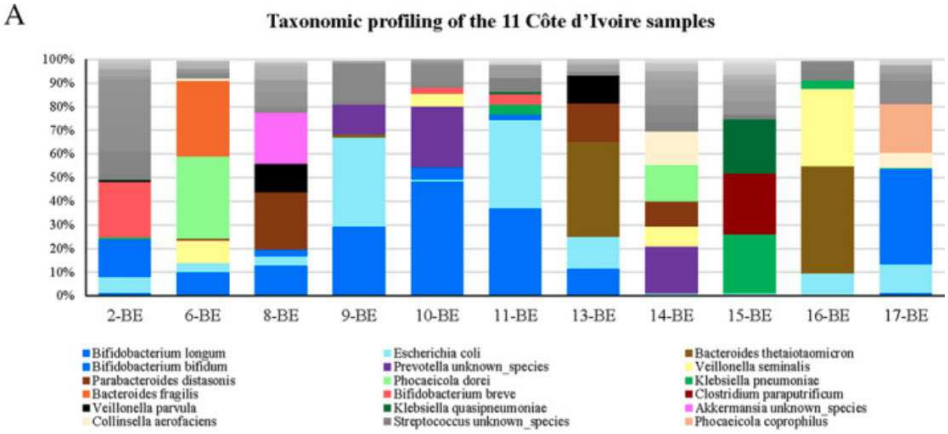
A preliminary 16S rRNA profiling analysis, encompassing the 11 Côte d'Ivoire samples, was performed to obtain an overview of the gut microbiota composition at the genus level. The 16S rRNA profiling is a technique based on sequencing of a specific region of the bacterial ribosomal locus, i.e. 16S region, which allows obtaining an accurate bacterial taxonomic classification down to genus level (Fig. S2; Table S2) (Costea et al., 2017; Mancabelli et al., 2020). The results, consistently with those described in the so-far published scientific literature, showed that the genera *Bacteroides*, *Bifidobacterium*, *Escherichia*, *Veillonella* and *Clostridium* were the dominant taxa in the 11 Côte d'Ivoire subjects, with an average relative abundance ranging from 7.42% to 18.63% (Table S2) (Decuypere et al., 2016; Laudadio et al., 2018).

Furthermore, prevalence analysis performed by considering microbial taxa with relative abundance >0.01% revealed that *Escherichia-Shigella* and *Streptococcus* were present with a prevalence of 100%. At the same time, members of the *Bifidobacterium* genus, with documented host' health benefits (Milani et al., 2017), were present in 10 samples from Côte d'Ivoire while absent only in one sample, named 15-BE (Table S2). This latter sample showed a high abundance of *Clostridium* and *Klebsiella* species, which are both correlated with an unhealthy status of the gut microbiota and remarkably this sample revealed the absence of *Bifidobacterium* (Table S2).

Species-level profiling through shallow shotgun metagenomics of Côte d'Ivoire infant gut microbiota.

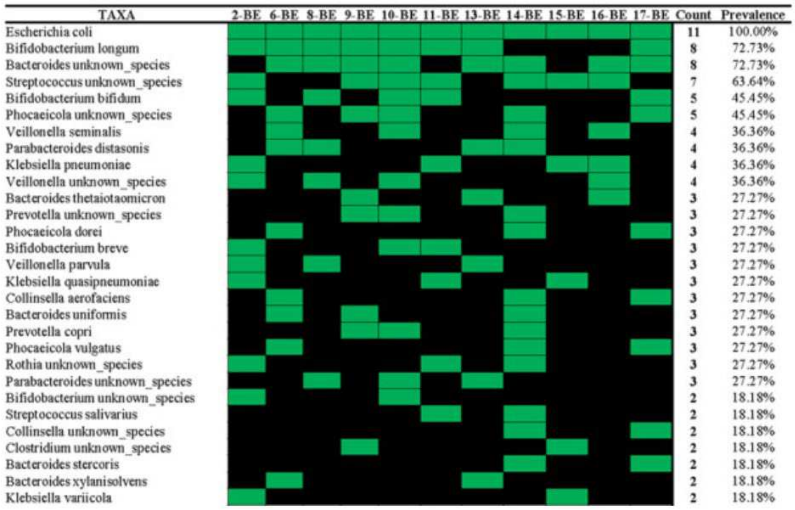
Following the assessment of the genera-level gut microbiota composition of the 11 Côte d'Ivoire samples, we re-analysed all samples through the shallow shotgun metagenomics method to achieve bacterial taxonomic classification at species-level (Hillmann et al., 2018). The above described 16S rRNA-based profiles at the genus level were compared with those obtained through shallow-shotgun metagenomics, showing a perfect match between them with high reliability (Fig. S2). Thus, shallow shotgun metagenomics analysis allowed us to properly decipher the intra-genera complexity of the faecal microbial communities (Table S3; Fig. 1) (Hillmann et al., 2018).

A



B

Prevalence Heatmap of the 11 Côte d'Ivoire samples



TAXA	2-BE	6-BE	8-BE	9-BE	10-BE	11-BE	13-BE	14-BE	15-BE	16-BE	17-BE	Count	Prevalence
Escherichia coli	■	■	■	■	■	■	■	■	■	■	■	11	100.00%
Bifidobacterium longum	■	■	■	■	■	■	■				■	8	72.73%
Bacteroides unknown_species		■	■	■	■			■	■	■	■	8	72.73%
Streptococcus unknown_species	■			■	■	■		■	■	■		7	63.64%
Bifidobacterium bifidum	■		■	■	■	■					■	5	45.45%
Phocaeicola unknown_species		■		■	■			■			■	5	45.45%
Veillonella seminalis		■			■			■		■		4	36.36%
Parabacteroides distasonis		■	■				■	■				4	36.36%
Klebsiella pneumoniae	■					■			■	■		4	36.36%
Veillonella unknown_species	■		■		■					■		4	36.36%
Bacteroides thetaiotaomicron				■			■			■		3	27.27%
Prevotella unknown_species				■	■			■				3	27.27%
Phocaeicola dorei		■						■			■	3	27.27%
Bifidobacterium breve	■				■	■						3	27.27%
Veillonella parvula	■		■				■					3	27.27%
Klebsiella quasipneumoniae	■					■			■			3	27.27%
Collinsella aerofaciens		■						■			■	3	27.27%
Bacteroides uniformis		■		■				■				3	27.27%
Prevotella copri				■	■			■				3	27.27%
Phocaeicola vulgatus		■						■			■	3	27.27%
Rothia unknown_species	■					■		■				3	27.27%
Parabacteroides unknown_species			■		■		■					3	27.27%
Bifidobacterium unknown_species	■				■							2	18.18%
Streptococcus salivarius						■		■				2	18.18%
Collinsella unknown_species								■			■	2	18.18%
Clostridium unknown_species				■					■			2	18.18%
Bacteroides stercoris								■			■	2	18.18%
Bacteroides xylanisolvens		■					■					2	18.18%
Klebsiella variicola	■								■			2	18.18%

Fig. 1. Taxonomic profiling of the 11 Côte d'Ivoire samples collected in this study. Panel A shows a bar plot representation of the taxonomic composition of the 11 Côte d'Ivoire infants' samples. Panel B reports a prevalence heatmap, showing taxa identified in at least two samples among the pool of 11 infants' samples.

Notably, the species-level prevalence matrix showed that 29 bacterial species are shared by two or more samples, covering more than 80% of the total average

compositions of each sample (Table S3; Fig. 1). Among these 29 species, three were shared by more than 40% of the samples (5 out of 11 samples at least), i.e. *Escherichia coli*, *Bifidobacterium longum* and *Bifidobacterium bifidum*, which also showed with high average abundance (between 6.05% and 13.88%) (Table S3; Fig. 1). Remarkably, these results confirm that members of the *Bifidobacterium* genus represent key early gut colonizers, along with *E. coli*, as previously observed for infants living in urbanized countries (Milani et al. 2017b).

Meta-analysis of species-level gut microbiota profiles of 1098 infants from publicly available datasets

A collection of 1098 publicly available shotgun samples, corresponding to faecal samples of urbanized and rural infants aged between a few days of life to 12 months, were retrieved (Table S1; Fig. S1). Upon the inclusion of the 11 Côte d'Ivoire shallow shotgun datasets, a total of 1109 samples were grouped by infants' age in 1–3 M category (n = 309, 0–3 months of life), 3–6 M category (n = 588, 3–6 months of life), 6–9 M category (n = 114, 6–9 months of life) and 9–12 M category (n = 133, 9–12 months of life) (Table S1). To prevent biases due to data analysis, all the shotgun metagenomics datasets were re-analysed using METAnnotatorX (Milani et al., 2018). In detail, all the datasets were quality-filtered and 100 000 quality-filtered reads were exploited for the species-level taxonomic reconstruction of the faecal microbiota composition through the shallow metagenomics approach (Hillmann et al., 2018).

The gut microbiota composition of the 1109 abovementioned metagenomics datasets was assessed (Table S4). Data retrieved showed that the most abundant

species was *B. longum*, with an average abundance of 16.07%, followed by *Escherichia coli*, *Bifidobacterium breve* and *B. bifidum* with an average abundance of 7.60%, 6.18% and 5.96% respectively (Table S4). Moreover, *B. longum*, *B. breve* and *E. coli* were present in more than 50% of the infants, showing a prevalence of 73.67%, 64.20% and 50.86% respectively (Table S5). Remarkably, the latter taxa were also the most prevalent and abundant in infants from Côte d'Ivoire sequenced in this study (Table S3; Fig. 1). Thus, suggesting that host-associated factors may be the main responsible in gut microbiota development, while environmental factors may participate in defining the wide range of interindividual variations.

Following the species-profiling analysis, in order to obtain a statistic-based clustering of the samples based on their taxonomic composition, we performed the hierarchical cluster analysis (Difference between K Means and Hierarchical Clustering - GeeksforGeeks, n.d.), leading to the generation of 20 clusters named from ISCSTs_1 to ISCSTs_20 with a clear and well-defined microbiological profile (Fig. 2; Table S6). Notably, the high number of predicted clusters compared to adult microbiota meta-analyses (Derrien et al., 2019) is compatible with the higher degree of interindividual variability during gut microbiota development (Milani et al., 2017; Derrien et al., 2019).

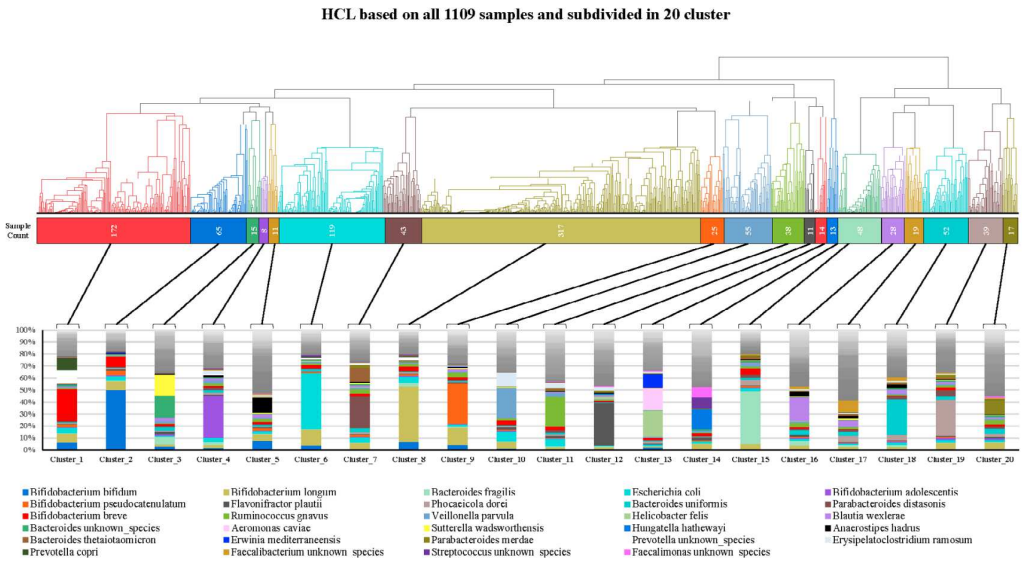


Fig. 2. Average taxonomic composition of the predicted ISCST. The representation reports the average taxonomic composition of the predicted ISCST along with the hierarchical clustering dendrogram based on the relative abundance table that was used to predict the ISCSTs.

Intriguingly, the three most prevalent clusters were ISCSTs_8, ISCSTs_1 and ISCSTs_6. In detail, 317 samples (28.58% of total) fall into the ISCSTs_8, showing a bacterial community profile with *B. longum* as dominant taxa (average abundance of 41.78%) (Table S6; Fig. 2). In contrast, 172 samples (15.51% of total samples) were included in ISCSTs_1, showing a co-dominance of *B. breve* (23.79% average abundance) and species belonging to the genus Prevotella, mainly *Prevotella copri* (9.02% average abundance) (Table S6; Fig. 2). Furthermore, ISCSTs_6, including 119 samples (10.73% of total samples), appeared to be the third most representative cluster, displaying a profile dominated mainly by the *Escherichia coli* species (average abundance of 40.72%) (Table S6; Fig. 2). Together, these results reinforced the notion that infants' gut microbiota is generally rich in the *Bifidobacterium* genus, followed

by E. coli as reported in previous literature describing the pivotal role of these early gut colonizers in healthy infants (Vemuri et al., 2018).

Indeed, among the dominant taxa of the other relatively minor ISCSTs we still found various species of bifidobacteria, such as *B. bifidum* in ISCSTs_2 and *Bifidobacterium pseudocatenulatum* in Cluster_9, as well as multiple species belonging to the *Bacteroides* genus, such as *Bacteroides fragilis* in ISCSTs_15 and *Bacteroides uniformis* in ISCSTs_18 (Fig. 2; Table S2).

Furthermore, analysis of beta-diversity depicted through 3D Principal Components analysis and correlation coefficient showed differences in the gut microbiota composition between the four age-based groups (PERMANOVA p-value <0.05) (Table S7; Fig. S3). This finding, coupled with those obtained from the alpha-diversity analysis, corroborates with the notion that the infants' gut microbiota experiences rapid changes during the first year of life (Factors Influencing the Gut Microbiome in Children: From Infancy to Childhood - PubMed, n.d.).

In this context, the 11 Côte d'Ivoire infants' samples, with age ranging from 1 to 90 days, enrolled in this study's framework showed average biodiversity of 14.54 species per sample, matching, as for age category 1–3 M.

Data regarding infant's age were also correlated to the 20 predicted clusters. Taking into account the age-based samples subdivision, we found that ISCSTs_16, ISCSTs_17 and ISCSTs_18 constituted primarily (between 55.77% and 82.15% average abundance) by older infants (6–9 M and 9–12 M categories), which were associated with higher alpha diversity (p-value <0.05) (Table S7), corresponding to an average of 23.9, 27.7 and 24.1 species per sample respectively (Figs S4 and S5). In contrast, ISCSTs_1, ISCSTs_2 and ISCSTs_6,

formed primarily (between 79.07% and 92.31% average abundance) by infants below 6 months of age (1–3 M and 3–6 M categories), showed an average of 16, 12.3, and 12.8 species per sample respectively. These findings are consistent with previous literature, demonstrating the progressive increase in gut microbiota biodiversity with infant' age (Milani et al., 2017; Derrien et al., 2019).

Species-level comparison of Côte d'Ivoire infant against ISCSTs clusters

The large pool of 1098 datasets from different continents allowed us to compare the worldwide taxonomic variability of the gut microbiota of newborns during the first year of life with that of 11 newborns living in the Côte d'Ivoire. In detail, the taxonomic profile of each of the 11 Côte d'Ivoire infants was compared to the average profile observed for each ISCST. Moreover, it was also compared with the average composition identified for the 275 samples available in public databases corresponding to infants living in rural regions of Africa to identify key microbial signatures of gut microbiota development in pre-urbanized countries.

Intriguingly, 2-BE, 17-BE, 9-BE, 8-BE, 13-BE, 16-BE, 10-BE and 11-BE, fall in ISCSTs represented primarily by samples of age categories A and B (pre-weaning), dominated by the bifidobacterial species such as *B*. *longum*, *B*. *bifidum*, *B*. *breve* and/or *E*. *coli*, i.e. ISCSTs 1, 2, 6, 7 and 8 (Figs 2 and and3).3). These data were in line with the age of the infants from which the samples were taken, which belonging to the 1–3 M category (<90 days of life) (Table S1; Fig. 2).

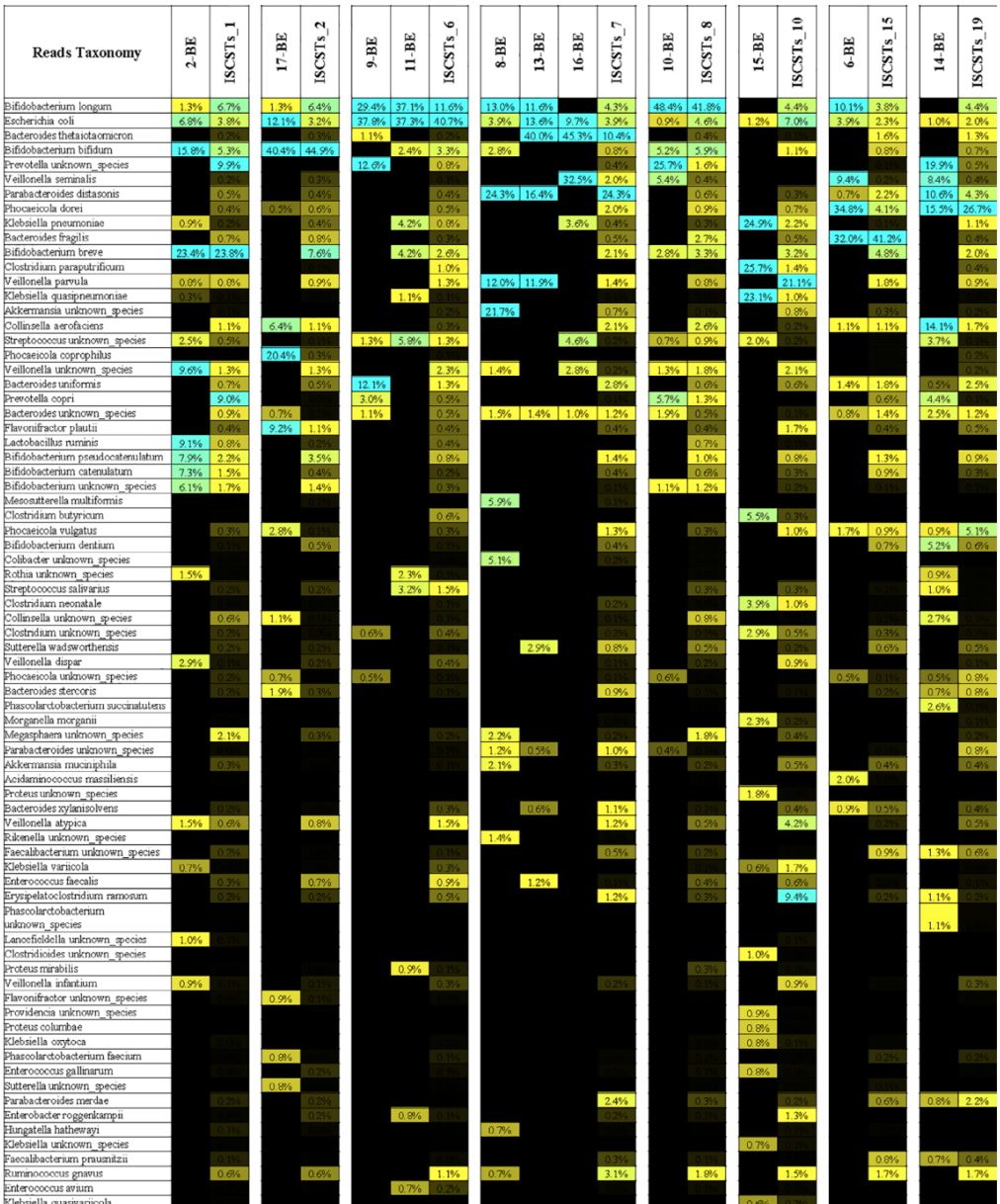


Fig. 3. Comparison between 11 Côte d'Ivoire samples and ISCSTs. A colour-graduated heatmap is reported in order to compare Côte d'Ivoire samples with predicted ISCSTs. Only taxa present in sub-Saharan pool samples are graphically reported.

In contrast, sample 6-BE was classified as a member of ISCST_15 due to the high relative abundance of *Bacteroides fragilis*. Remarkably, 28.83% of ISCST_15 is constituted by datasets of individuals of age category 9–12 M (>270 days), i.e. post-weaning, as expected due to high relative abundance of *Bacteroides* species. These data may indicate the habit, in developing contexts, of introducing solid food early in the diet, especially concerning 6-BE, a mixed feeding infant (Table S1).

The remaining samples 14-BE and 15-BE showed a high relative abundance of *Phocaeicola dorei* and *C. paraputrificum* along with various species of *Klebsiella* respectively and were assigned to ISCST_19 and ISCST_10, (Figs 2 and and3;3; Table S6). Intriguingly, *C. paraputrificum* has been associated with paediatric infection, bacteremia, adult sepsis, and could indicate a risk for gut infections for 14-BE and 15-BE (Kiu et al., 2017; Intra et al., 2020).

The 11 samples of infants living in semi-urban African context sequenced in this study were also compared to a dataset of 275 infants living in Africa with limited interaction with urbanized populations that were released in the framework of previous studies (Table S1) (Bender et al., 2016; Agapova et al., 2018). The latter datasets span from 1–3 M to 3–6 M age and are distributed mainly in three clusters associated with the same age range, i.e. ISCST_1, 6 and 8, due to dominance of three main taxa, *B. longum*, *E. coli* and *B. breve* (Table 1; Fig. 2). Notably, only 3.64% of the samples constituting this dataset fall in ISCSTs_2, 7 and 10, which are also associated with age groups 1–3 M/3–6 M and are highly represented in urbanized populations (17.89% of samples). This limited diversity in terms of ISCSTs observed for pre-urbanized infants is probably correlated with a dietary intake in rural communities that causes the selection of a limited number

of dominant species, especially compared to the much more varied diets of the urban infants (Voreades et al., 2014; Savage et al., 2018). Diet has been observed to influence the human milk oligosaccharides (HMO) profiles of lactating mothers. Thus, putatively affecting the selection of bacterial taxa in newborns by modulating the HMO profiles (Seferovic et al., 2020).

Table 1

Urban, rural and sub-Saharan samples correlation with ISCSTs.

	Urban		Rural		11 Sub-urban	
ISCSTs	**Number of samples**	**%**	**Number of samples**	**%**	**Number of samples**	**%**
1	105	12.76	66	24.00	1	9.09
2	61	7.41	3	1.09	1	9.09
3	15	1.82	0	0.00	0	0.00
4	7	0.85	1	0.36	0	0.00
5	11	1.34	0	0.00	0	0.00
6	73	8.87	44	16.00	2	18.18
7	37	4.50	3	1.09	3	27.27
8	173	21.02	143	52.00	1	9.09
9	19	2.31	6	2.18	0	0.00
10	50	6.08	4	1.45	1	9.09
11	36	4.37	2	0.73	0	0.00
12	11	1.34	0	0.00	0	0.00
13	14	1.70	0	0.00	0	0.00
14	12	1.46	1	0.36	0	0.00
15	45	5.47	2	0.73	1	9.09
16	28	3.40	0	0.00	0	0.00
17	19	2.31	0	0.00	0	0.00
18	52	6.32	0	0.00	0	0.00
19	38	4.62	0	0.00	1	9.09
20	17	2.07	0	0.00	0	0.00
Total samples	823		275		11	

Notably, in contrast to the 11 infant fecal samples from a semi-urban area in Côte d'Ivoire, the 275 samples from two African rural communities showed a higher relative abundance of *B. longum* (t-test p-value <0.001) (Table_S7).

Covariance network analysis based on compositions of 1109 infants' samples.

To define the intricate positive and negative relationships existing between the bacterial species constituting the 1109 infants' faecal samples, we performed a bivariate covariance analysis based on the Pearson correlations with a bootstrap value of 1000 repetitions (Table S8) (Software SPSS - Italia | IBM, n.d.). To remove background noise, we selected the bacterial species present in more than 10 samples. The data resulting from the covariance analysis were exploited to create an interaction network with Gephy software (Gephi - The Open Graph Viz Platform, n.d.) with a modularity value set to 1.5 (Fig. 4; Table S9) (Brandes et al., 2008). This analysis allowed us to identify four Modularity Cluster (MC), named MC_1, MC_2, MC_3 and MC_4, in which the species forming a given cluster tend to correlate positively with each other and negatively with the members of other clusters.

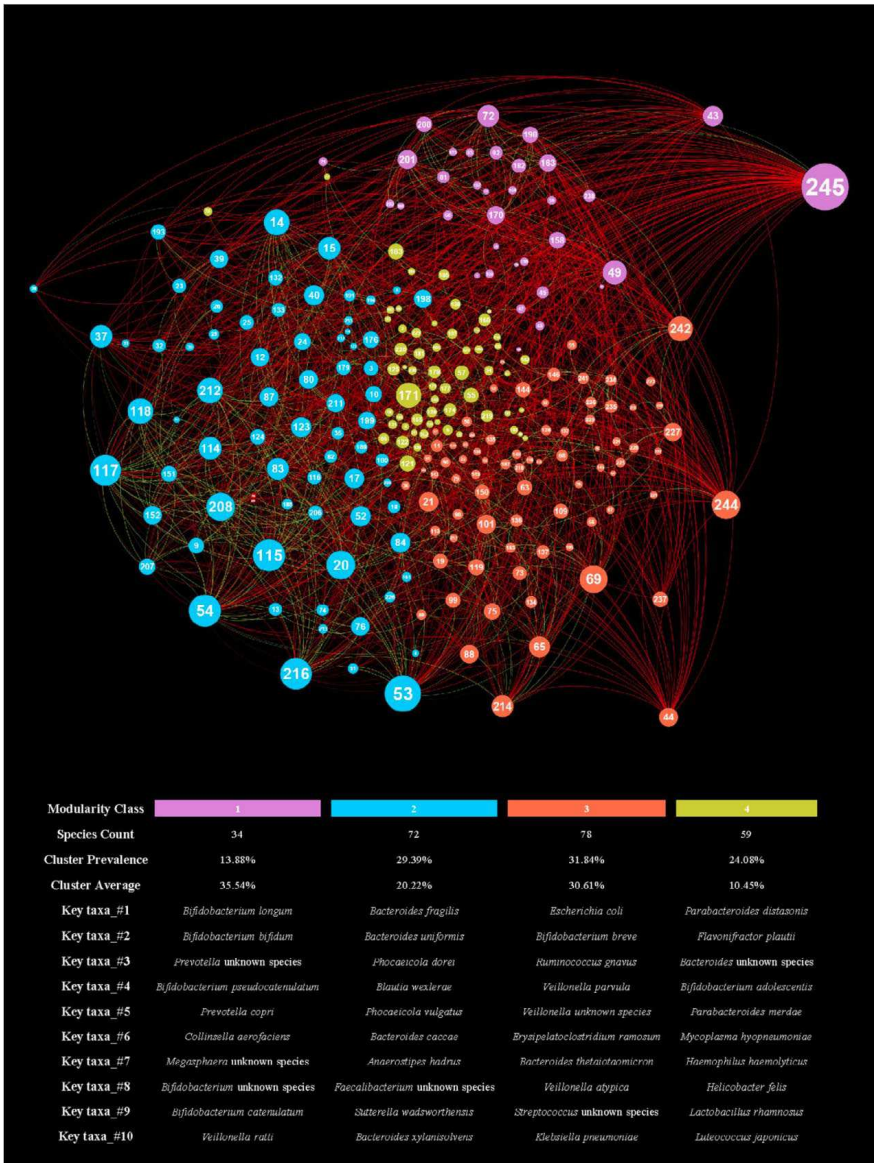


Fig. 4. Network covariance of taxa observed in the 1109 samples included in the meta-analysis. A network generated via Gephi software and force atlas 2 algorithm is reported in order to graphically represent the covariance relationship between each taxa observed in at least 10 samples. This filtering was made to remove background noise and enhance the clearness of the image.

In detail, MC_1 is mainly constituted by health-promoting taxa typically associated with the early gut microbiota development, such as members of the *Prevotella* and Bifidobacterium genera, including B. longum, B. bifidum and *P. copri*. Interestingly, although a limited range of microbial taxa encompassed this cluster (13.88% of the analysed species), they corresponded to a sum of 35.54% of average abundance (Fig. 4; Table S9). Moreover, MC_2 is mainly composed of species belonging to *Bacteroides*, *Phocaeicola* and *Faecalibacterium* genera, such as *Bacteroides fragilis*, *P. dorei* and *Faecalibacterium prausnitzii*, which have been previously correlated with a healthy adult gut microbiota environment. This cluster covers 29.39% of all the analysed species, corresponding to a total summed average abundance of 20.22% (Fig. 3; Table S9). Intriguingly, these two clusters highlight that a specific set of bacterial taxa seem to cooperate for the subsequent colonization of the two peculiar ecological niches represented by the gut environment of pre-weaning (MC_1) and post-weaning (MC_2) infants.

Furthermore, MC_3 is composed primarily of species belonging to *Escherichia*, *Veillonella*, *Klebsiella* and *Clostridium* genera, covering 31.84% of all the analysed species, corresponding to a sum of an average abundance of 30.61% (Fig. 4; Table S9). This cluster highlighted how potentially health-threatening species, such as *E. coli* and *Clostridium*, tend to correlate positively with each other. However, this does not mean that they can directly cause harm to the host's health (Christofi et al., 2019). Members of these taxa are commonly found in the healthy intestinal microbiota and are also engaged in positive interactions with probiotic species such as *B. breve* (Cukrowska et al., 2020). Remarkably, the latter result supports the notion that species belonging to *Escherichia*, *Veillonella*, *Klebsiella* and *Clostridium* genera may act as opportunistic

pathogens, i.e. may negatively affect the host's health only in specific conditions of altered gut microbiota homeostasis.

Lastly, MC_4 is a cluster composed primarily of accessory taxa, including *Parabacteroides distasonis*, *Flavonifractor plautii* and *Bifidobacterium adolescentis*, which showed weak positive interactions with the other dominant clusters. This cluster covers 24.08% of all analysed species, corresponding to a sum of average abundances of only 10.45% (Fig. 4; Table S9). Notably, this cluster encompassed a wide range of bacterial species typically found across all ages as minor players of the gut community. Thus, suggesting that these bacterial taxa, despite their possible cooperation for efficient niche colonization, tend to limitedly interact with members of other MC clusters and may exert a key role in defining inter-individual variability in the gut microbiota composition (Almeida et al., 2019; Yang et al., 2020).

By correlating the compositions of the 11 Côte d'Ivoire infants with the MCs, we observed that most of their taxonomic composition falls into MC_3 and MC_1 apart from the 6-BE sample, which has >70% of its taxonomic composition belonging to MC_2 (Fig. S6; Table S9). Intriguingly, only 15-BE and 16-BE have >90% of their taxonomic composition belonging to MC_3, while showing limited participation of MC_1. This could indicate a possible onset of future pathogenic dysbiotic states due to the marked dominance of opportunistic pathogens (Fig. S6; Table S9).

Furthermore, by correlating rural, semi-urban and urban samples with MCs, we revealed that in rural datasets the species belonging to MC_1 make up for 66.26% of the total taxonomic composition, while MC_1 covers only 25.31% in urban datasets. In rural datasets, an additional 25.16% of the total taxonomic

composition belongs to MC_3. At the same time, urban samples are characterized by higher variability, with 25.64% of the taxonomic composition belonging to MC_2, 32.23% to the MC_3 cluster and 12.96% to MC_4 (Fig. S6).

Conclusions

The large pool of shotgun metagenomics datasets corresponding to 1098 infants included in this meta-analysis allowed us to explore the worldwide taxonomic variability of the infant gut microbiota across the first year of life. Furthermore, in the framework of this study, we collected a faecal sample from 11 infants living in Côte d'Ivoire. The latter, along with two additional publicly available datasets corresponding to infants living in pre-urbanized populations, revealed microbial signatures associated with pre-urbanization. Intriguingly, data collected suggests that the geographical origin and diet of pre-urbanized populations impact the overall microbial biodiversity of the infant gut microbiota. In fact, mothers' diet in pre-urbanized areas, defined by high consumption of simple sugars, seasonal foods and fibre intake, can indirectly determine differences in infant's gut microbiota through the modulation of HMO's profiles of lactating mothers (Seferovic et al., 2020). Indeed, modulation of the prevalence of the main enterotypes, i.e. ISCSTs, was predicted by statistical integration of all the taxonomic profiles retrieved by the meta-analysis. Remarkably, while this study revealed intriguing data, further investigations with additional larger cohorts are needed to validate and extend our knowledge of the gut microbiota development of infants living in pre-urbanized areas.

Author contribution

R.D.A conceived the study, performed samples collections and wrote the manuscript. F.F. performed bioinformatic analyses and wrote the manuscript. C.M performed bioinformatic analyses, conceived the study, and wrote the manuscript. L.M, G.A.L., C.T. performed data analysis. R.A., G.L., G.A. and A.V. performed microbial DNA sequencing. M.V. and F.T. revised and approved the manuscript. R.F., M.B., K.H.M., performed samples collections. All authors reviewed the manuscript.

Acknowledgments

We thank Mr. Ayémou Antoine Ago and the Clinical Laboratory of the Centre Médical Don Orione of Anyama, for their very helpful collaboration. We furthermore thank GenProbio srl for financial support of the Laboratory of Probiogenomics. This research benefited from the HPC (High Performance Computing) facility of the University of Parma, Italy.

Data availability

The SRA accession numbers of the metagenomic sequences of the 11 Côte d'Ivoire infant faecal sample sequenced in this study is SRP311268.

Funding

RD research was funded by Mission Sustainability-2017 program n. E86C18000570005, University of Rome Tor Vergata.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1: Supplementary File.

Table S1. Metadata of samples included in this meta-analysis.

Table S2. 16S rRNA sequencing-based taxonomic profiling of the 11 Côte d'Ivoire samples.

Table S3. Shallow shotgun-based taxonomic profiling of the 11 Côte d'Ivoire samples.

Table S4. Shallow shotgun-based taxonomic profiling of the 1109 samples included in this meta-analysis.

Table S5. Full heatmap presence-absence of all the bacterial species detected in the 1109 samples included in this meta-analysis.

Table S6. Average ISCSTs taxonomic composition at species-level.

Table S7. Table of all the statistical analyses performed in this meta-analysis.

Table S8. Pearson-based matrix of the covariances observed between bacterial taxa predicted in at least 10 samples included in the meta-analysis.

Table S9. Taxonomic composition of the predicted modularity clusters.

Fig. S1. Metadata of the 1109 samples included in this meta-analysis. In panel a is reported a cake graph explaining geographic subdivision of the 1109 samples. In panel b is reported a cake graph showing age subdivision of the 1109 samples.

Fig. S2. Graphic comparison between 16S rRNA gene microbial profiling and shallow metagenomic profiling at genus level. In panel a a bar plot is reported in order to show average abundance compositions at genera level retrieved through

shallow shotgun profiling. In panel b a bar plot is displayed in order to show average abundance compositions at genera level obtained by 16S rRNA gene microbial profiling, and only taxa >0.1% Average are showed for cleanness.

Fig. S3. Beta-diversity analysis of the 1109 samples included in the meta-analysis. Panels a and b show a PCoA representation based on the Bray-Curtis index and the species-level taxonomic profile obtained for the 1109 samples included in the meta-analysis. The samples are colored based on age groups in panel a, and based on ISCST in panel b.

Fig. S4. ISCSTs age compositions. In panel a is shown a bar plot representation of the age group composition of every ISCSTs as sample counts. In panel b is reported a bar plot representation of the age group composition of every ISCSTs as percentage of the whole ISCST. Panel c provides a detailed summary of all the metadata associated with the predicted ISCSTs.

Fig. S5. Average alpha diversity of the predicted ISCSTs. In panel a is reported a bar plot representation of the raw count of the number of species identified in each ISCSTs. Panel b shows a bar plot representing the average number of species (Alpha diversity) correlated to each ISCSTs.

Fig. S6. Modularity clusters correlated to the 11 sub-Saharan samples. In panel a is reported a bar plot representation of sub-Saharan sample composition in terms of previously defined MCs. Panel b shows a table detailed data regarding composition in terms of previously predicted MCs.

References

Adak, Atanu, and Mojibur R. Khan. 2019. “An Insight into Gut Microbiota and Its Functionalities.” *Cellular and Molecular Life Sciences*. Birkhauser Verlag AG. https://doi.org/10.1007/s00018-018-2943-4.

Agapova, Sophia E., Kevin B. Stephenson, Oscar Divala, Yankho Kaimila, Kenneth M. Maleta, Chrissie Thakwalakwa, M. Isabel Ordiz, Indi Trehan, and Mark J. Manary. 2018. “Additional Common Bean in the Diet of Malawian Children Does Not Affect Linear Growth, but Reduces Intestinal Permeability.” *Journal of Nutrition* 148 (2): 267–74. https://doi.org/10.1093/jn/nxx013.

Almeida, Alexandre, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley, and Robert D. Finn. 2019. “A New Genomic Blueprint of the Human Gut Microbiota.” *Nature* 568 (7753): 499–504. https://doi.org/10.1038/s41586-019-0965-1.

Bender, Jeffrey M., Fan Li, Shoria Martelly, Erin Byrt, Vanessa Rouzier, Marguerite Leo, Nicole Tobin, et al. 2016. “Maternal HIV Infection Influences the Microbiome of HIV-Uninfected Infants.” *Science Translational Medicine* 8 (349): 349ra100-349ra100. https://doi.org/10.1126/scitranslmed.aaf5103.

Bolyen, Evan, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, et al. 2019. “Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2.” *Nature Biotechnology*. Nature Publishing Group. https://doi.org/10.1038/s41587-019-0209-9.

Brandes, Ulrik, Daniel Delling, Marco Gaertler, Robert Görke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. n.d. “On Modularity Clustering.”

Christofi, Theodoulakis, Stavria Panayidou, Irini Dieronitou, Christina Michael, and Yiorgos Apidianakis. 2019. “Metabolic Output Defines Escherichia Coli as a Health-Promoting Microbe against Intestinal Pseudomonas Aeruginosa.” *Scientific Reports* 9 (1). https://doi.org/10.1038/s41598-019-51058-3.

Costea, Paul I., Falk Hildebrand, Arumugam Manimozhiyan, Fredrik Bäckhed, Martin J. Blaser, Frederic D. Bushman, Willem M. De Vos, et al. 2017. “Enterotypes in the Landscape of Gut Microbial Community Composition.” *Nature Microbiology* 3 (1): 8–16. https://doi.org/10.1038/s41564-017-0072-8.

Cukrowska, Bożena, Joanna B. Bierła, Magdalena Zakrzewska, Mark Klukowski, and Elżbieta Maciorkowska. 2020a. “The Relationship between the Infant Gut Microbiota and Allergy. The Role of Bifidobacterium Breve and Prebiotic Oligosaccharides in the Activation of Anti-Allergic Mechanisms in Early Life.” *Nutrients*. MDPI AG. https://doi.org/10.3390/nu12040946.

Decuypere, Saskia, Conor J. Meehan, Sandra Van Puyvelde, Tessa De Block, Jessica Maltha, Lompo Palpouguini, Marc Tahita, Halidou Tinto, Jan Jacobs, and Stijn Deborggraeve. 2016. "Diagnosis of Bacterial Bloodstream Infections: A 16S Metagenomics Approach." *PLoS Neglected Tropical Diseases* 10 (2). https://doi.org/10.1371/journal.pntd.0004470.

Derrien, Muriel, Anne Sophie Alvarez, and Willem M. de Vos. 2019. "The Gut Microbiota in the First Decade of Life." *Trends in Microbiology*. Elsevier Ltd. https://doi.org/10.1016/j.tim.2019.08.001.

"Difference between K Means and Hierarchical Clustering - GeeksforGeeks." n.d. Accessed February 11, 2021. https://www.geeksforgeeks.org/difference-between-k-means-and-hierarchical-clustering/.

Doare, Kirsty Le, Beth Holder, Aisha Bassett, and Pia S. Pannaraj. 2018. "Mother's Milk: A Purposeful Contribution to the Development of the Infant Microbiota and Immunity." *Frontiers in Immunology*. Frontiers Media S.A. https://doi.org/10.3389/fimmu.2018.00361.

"Factors Influencing the Gut Microbiome in Children: From Infancy to Childhood - PubMed." n.d. Accessed February 26, 2021. https://pubmed.ncbi.nlm.nih.gov/31180062/.

Gentile, Christopher L., and Tiffany L. Weir. 2018. "The Gut Microbiota at the Intersection of Diet and Human Health." *Science*. American Association for the Advancement of Science. https://doi.org/10.1126/science.aau5812.

"Gephi - The Open Graph Viz Platform." n.d. Accessed February 25, 2021. https://gephi.org/.

Hillmann, Benjamin, Gabriel A. Al-Ghalith, Robin R. Shields-Cutler, Qiyun Zhu, Daryl M. Gohl, Kenneth B. Beckman, Rob Knight, and Dan Knights. 2018a. "Evaluating the Information Content of Shallow Shotgun Metagenomics." *MSystems* 3 (6). https://doi.org/10.1128/msystems.00069-18.

Intra, J., A. Milano, C. Sarto, and P. Brambilla. 2020. "A Rare Case of Clostridium Paraputrificum Bacteremia in a 78-Year-Old Caucasian Man Diagnosed with an Intestinal Neoplasm." *Anaerobe* 66 (December). https://doi.org/10.1016/j.anaerobe.2020.102292.

Kiu, Raymond, Shabhonam Caim, Cristina Alcon-Giner, Gusztav Belteki, Paul Clarke, Derek Pickard, Gordon Dougan, and Lindsay J. Hall. 2017. "Preterm Infant-Associated Clostridium Tertium, Clostridium Cadaveris, and Clostridium Paraputrificum Strains: Genomic and Evolutionary Insights." *Genome Biology and Evolution* 9 (10): 2707–14. https://doi.org/10.1093/gbe/evx210.

Laudadio, Ilaria, Valerio Fulci, Francesca Palone, Laura Stronati, Salvatore Cucchiara, and Claudia Carissimi. 2018. "Quantitative Assessment of Shotgun Metagenomics and 16S RDNA Amplicon Sequencing in the Study of Human Gut Microbiome." *OMICS A Journal of Integrative Biology* 22 (4): 248–54. https://doi.org/10.1089/omi.2018.0013.

Leoz, Maria Lorna A. De, Karen M. Kalanetra, Nicholas A. Bokulich, John S. Strum, Mark A. Underwood, J. Bruce German, David A. Mills, and Carlito B. Lebrilla. 2015. "Human Milk

Glycomics and Gut Microbial Genomics in Infant Feces Show a Correlation between Human Milk Oligosaccharides and Gut Microbiota: A Proof-of-Concept Study." *Journal of Proteome Research* 14 (1): 491–502. https://doi.org/10.1021/pr500759e.

Lugli, Gabriele Andrea, Sabrina Duranti, Christian Milani, Leonardo Mancabelli, Francesca Turroni, Giulia Alessandri, Giulia Longhi, et al. 2020. "Investigating Bifidobacteria and Human Milk Oligosaccharide Composition of Lactating Mothers." *FEMS Microbiology Ecology* 96 (5). https://doi.org/10.1093/femsec/fiaa049.

Mancabelli, Leonardo, Chiara Tarracchini, Christian Milani, Gabriele Andrea Lugli, Federico Fontana, Francesca Turroni, Douwe van Sinderen, and Marco Ventura. 2020. "Multi-Population Cohort Meta-Analysis of Human Intestinal Microbiota in Early Life Reveals the Existence of Infant Community State Types (ICSTs)." *Computational and Structural Biotechnology Journal* 18 (January): 2480–93. https://doi.org/10.1016/j.csbj.2020.08.028.

Marchesi, Julian R., David H. Adams, Francesca Fava, Gerben D.A. Hermes, Gideon M. Hirschfield, Georgina Hold, Mohammed Nabil Quraishi, et al. 2016. "The Gut Microbiota and Host Health: A New Clinical Frontier." *Gut* 65 (2): 330–39. https://doi.org/10.1136/gutjnl-2015-309990.

Matamoros, Sebastien, Christele Gras-Leguen, Françoise Le Vacon, Gilles Potel, and Marie France De La Cochetiere. 2013. "Development of Intestinal Microbiota in Infants and Its Impact on Health." *Trends in Microbiology*. Trends Microbiol. https://doi.org/10.1016/j.tim.2012.12.001.

Milani, Christian, Eoghan Casey, Gabriele Andrea Lugli, Rebecca Moore, Joanna Kaczorowska, Conor Feehily, Marta Mangifesta, et al. 2018. "Tracing Mother-Infant Transmission of Bacteriophages by Means of a Novel Analytical Tool for Shotgun Metagenomic Datasets: METAnnotatorX." *Microbiome* 6 (1). https://doi.org/10.1186/s40168-018-0527-z.

Milani, Christian, Sabrina Duranti, Francesca Bottacini, Eoghan Casey, Francesca Turroni, Jennifer Mahony, Clara Belzer, et al. 2017a. "The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota." *Microbiology and Molecular Biology Reviews* 81 (4). https://doi.org/10.1128/mmbr.00036-17.

———. 2017b. "The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota." *Microbiology and Molecular Biology Reviews* 81 (4). https://doi.org/10.1128/mmbr.00036-17.

Nagpal, Ravinder, and Yuichiro Yamashiro. 2018. "Gut Microbiota Composition in Healthy Japanese Infants and Young Adults Born by C-Section." *Annals of Nutrition and Metabolism*. S. Karger AG. https://doi.org/10.1159/000490841.

"Origin 2021 Feature Highlights." n.d. Accessed January 22, 2021. https://www.originlab.com/2021.

Rutayisire, Erigene, Kun Huang, Yehao Liu, and Fangbiao Tao. 2016. “The Mode of Delivery Affects the Diversity and Colonization Pattern of the Gut Microbiota during the First Year of Infants’ Life: A Systematic Review.” *BMC Gastroenterology* 16 (1). https://doi.org/10.1186/s12876-016-0498-0.

Savage, Jessica H., Kathleen A. Lee-Sarwar, Joanne E. Sordillo, Nancy E. Lange, Yanjiao Zhou, George T. O’Connor, Megan Sandel, et al. 2018. “Diet during Pregnancy and Infancy and the Infant Intestinal Microbiome.” *Journal of Pediatrics* 203 (December): 47-54.e4. https://doi.org/10.1016/j.jpeds.2018.07.066.

Savino, F, F Cresi, S Pautasso, E Palumeri, V Tullio, J Roana, L Silvestro, and R Oggero. 2004. “Intestinal Microflora in Breastfed Colicky and Non-Colicky Infants.” *Acta Paediatrica* 93 (6): 825–29. https://doi.org/10.1111/j.1651-2227.2004.tb03025.x.

Schirbel, Anja, and Claudio Fiocchi. 2011. “Targeting the Innate Immune System in Pediatric Inflammatory Bowel Disease.” *Expert Review of Gastroenterology and Hepatology* 5 (1): 33–41. https://doi.org/10.1586/egh.10.76.

Seferovic, Maxim D., Mahmoud Mohammad, Ryan M. Pace, Melinda Engevik, James Versalovic, Lars Bode, Morey Haymond, and Kjersti M. Aagaard. 2020. “Maternal Diet Alters Human Milk Oligosaccharide Composition with Implications for the Milk Metagenome.” *Scientific Reports* 10 (1). https://doi.org/10.1038/s41598-020-79022-6.

“Software SPSS - Italia | IBM.” n.d. Accessed February 25, 2021. https://www.ibm.com/it-it/analytics/spss-statistics-software.

Turroni, Francesca, Christian Milani, Sabrina Duranti, Gabriele Andrea Lugli, Sergio Bernasconi, Abelardo Margolles, Francesco Di Pierro, Douwe Van Sinderen, and Marco Ventura. 2020. “The Infant Gut Microbiome as a Microbial Organ Influencing Host Well-Being.” *Italian Journal of Pediatrics*. BioMed Central Ltd. https://doi.org/10.1186/s13052-020-0781-0.

Vemuri, Ravichandra, Rohit Gundamaraju, Madhur D. Shastri, Shakti Dhar Shukla, Krishnakumar Kalpurath, Madeleine Ball, Stephen Tristram, Esaki M. Shankar, Kiran Ahuja, and Rajaraman Eri. 2018. “Gut Microbial Changes, Interactions, and Their Implications on Human Lifecycle: An Ageing Perspective.” *BioMed Research International*. Hindawi Limited. https://doi.org/10.1155/2018/4178607.

Voreades, Noah, Anne Kozil, and Tiffany L. Weir. 2014. “Diet and the Development of the Human Intestinal Microbiome.” *Frontiers in Microbiology* 5 (SEP). https://doi.org/10.3389/fmicb.2014.00494.

Yadav, Monika, Manoj Kumar Verma, and Nar Singh Chauhan. 2018. “A Review of Metabolic Potential of Human Gut Microbiome in Human Nutrition.” *Archives of Microbiology*. Springer Verlag.

https://doi.org/10.1007/s00203-017-1459-x.

Yang, Jing, Ji Pu, Shan Lu, Xiangning Bai, Yangfeng Wu, Dong Jin, Yanpeng Cheng, et al. 2020. "Species-Level Analysis of Human Gut Microbiota With Metataxonomics." *Frontiers in Microbiology* 11 (August): 2029. https://doi.org/10.3389/FMICB.2020.02029.

Chapter 6

Multifactorial Microvariability of the Italian Raw Milk Cheese Microbiota and Implication for Current Regulatory Scheme

Federico Fontana1,2#, Giulia Longhi1,2#, Giulia Alessandri1, Gabriele Andrea Lugli1, Leonardo Mancabelli1,3, Chiara Tarracchini1, Alice Viappiani2, Rosaria Anzalone2, Marco Ventura1,3, Francesca Turroni1,3* and Christian Milani1,3*

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

#These authors contributed equally.
*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 specifications. 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.

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 that follows 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 harboring 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 production site-specific high variability are only 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 preactivated 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 microgeographical scale such as temperature and humidity levels, ultimately causing fluctuations in the final organoleptic features of the dairy product.

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) (Data Set S1).

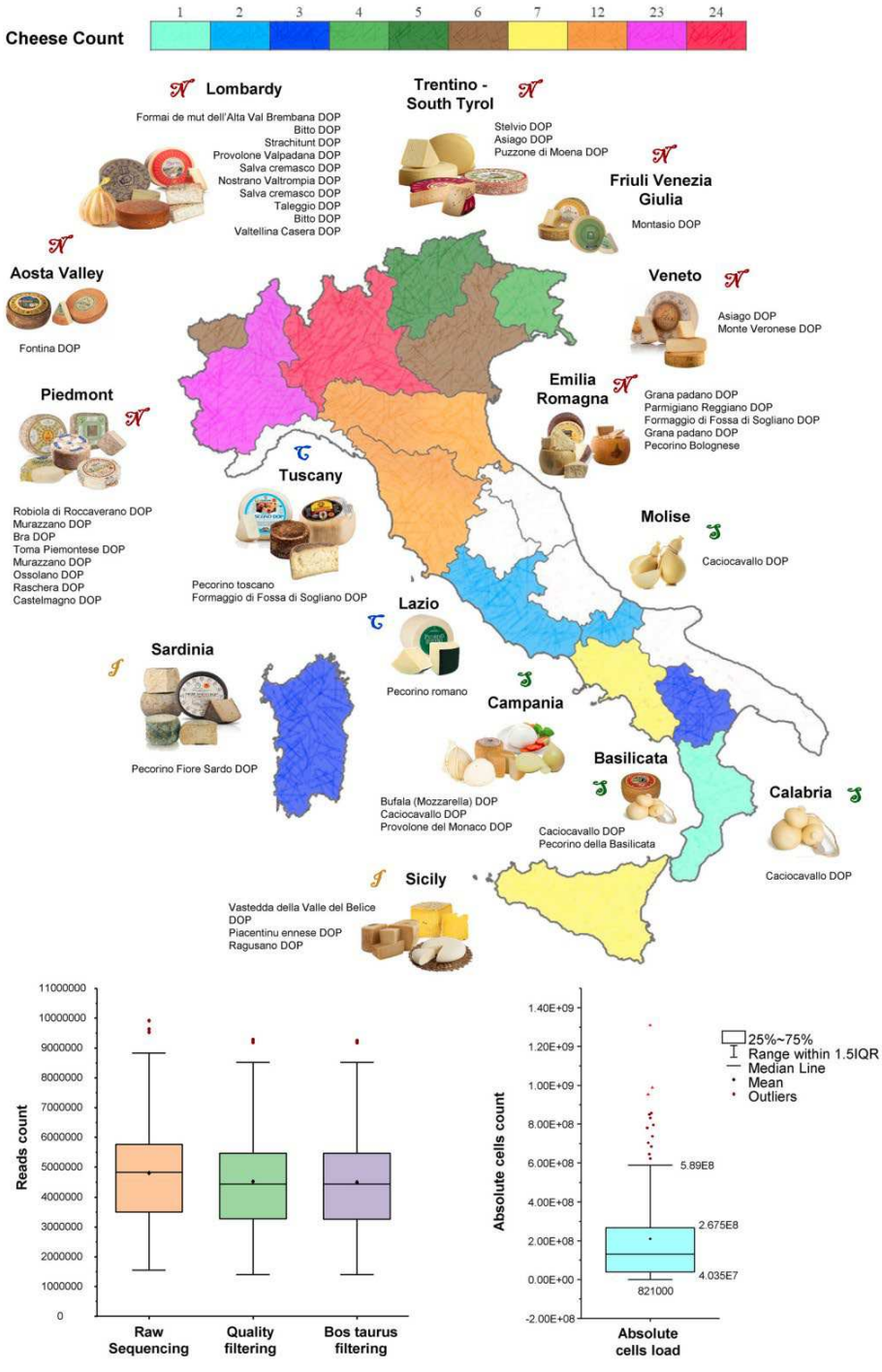

Cheese Count
1
2
3
4
5
6
7
12
23
24
Lombardy
Formai de mut dell'Alta Val Brembana DOP
Bitto DOP
Strachitunt DOP
Provolone Valpadana DOP
Salva cremasco DOP
Nostrano Valtrompia DOP
Salva cremasco DOP
Taleggio DOP
Bitto DOP
Valtellina Casera DOP
Trentino - South Tyrol
Stelvio DOP
Asiago DOP
Puzzone di Moena DOP
Friuli Venezia Giulia
Montasio DOP
Aosta Valley
Fontina DOP
Veneto
Asiago DOP
Monte Veronese DOP
Piedmont
Robiola di Roccaverano DOP
Murazzano DOP
Bra DOP
Toma Piemontese DOP
Murazzano DOP
Ossolano DOP
Raschera DOP
Castelmagno DOP
Emilia Romagna
Grana padano DOP
Parmigiano Reggiano DOP
Formaggio di Fossa di Sogliano DOP
Grana padano DOP
Pecorino Bolognese
Tuscany
Pecorino toscano
Formaggio di Fossa di Sogliano DOP
Molise
Caciocavallo DOP
Lazio
Pecorino romano
Sardinia
Pecorino Fiore Sardo DOP
Campania
Bufala (Mozzarella) DOP
Caciocavallo DOP
Provolone del Monaco DOP
Basilicata
Caciocavallo DOP
Pecorino della Basilicata
Calabria
Caciocavallo DOP
Sicily
Vastedda della Valle del Belice DOP
Piacentinu ennese DOP
Ragusano DOP
Reads count
11000000
10000000
9000000
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
Raw Sequencing
Quality filtering
Bos taurus filtering
Absolute cells count
1.40E+09
1.20E+09
1.00E+09
8.00E+08
6.00E+08
4.00E+08
2.00E+08
0.00E+00
-2.00E+08
25%~75%
Range within 1.5IQR
Median Line
Mean
Outliers
5.89E8
2.675E8
4.035E7
821000
Absolute cells load

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

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 (Data Set S1) (Fig. 1). Subsequently, a flow cytometry assay of the total bacterial load present in 0.2 g of cheese was used to transform the relative abundance of each profiled microbial taxa into absolute abundance, i.e., estimation of species-specific cells loads (Data Set S1) (Fig. 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) (Data Set S1).

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%) (Data Set S1).

Notably, despite a core microbiota consisting of 11 highly prevalent species, visualization of the intersample's taxonomic diversity (beta-diversity) through a two-dimensional principal coordinate analysis (PCoA) revealed the absence of evident clustering of cheeses based on cheese type or regional localization (Fig. S1). Nevertheless, validation through ANOSIM analysis revealed an R correlation of 12.8% ($P < 0.005$) (Fig. S1) indicative that geographical region partially participates in defining the taxonomic composition. In-depth statistical investigation (detailed in the Text S1) 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 (Data Set S1). In contrast, no correlation between geographical region and cheese microbiota was found for PDO cheeses (Data Set S1).

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 (18, 22). 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%, $P < 0.001$) (Fig. S3). Then, through a PCoA analysis, we noticed that most cheeses with hard matrices tend to cluster together. In contrast, semihard and soft cheeses did not show any particular clustering profile (Fig. S3). In detail, between the hard cheeses only

two PDO types seem to cluster together, i.e., Parmigiano Reggiano and Grana Padano (Fig. S1, Fig. S2 and S3). 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, nonaged 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 (Data Set S1) (Fig. S1 and S3).

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 (Data Set S1) (Fig. 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) (Data Set S1). The average bacterial load observed for the predicted HPCCSTs ranged from 6.14E + 07 to 2.44E + 08 (Data Set S1).

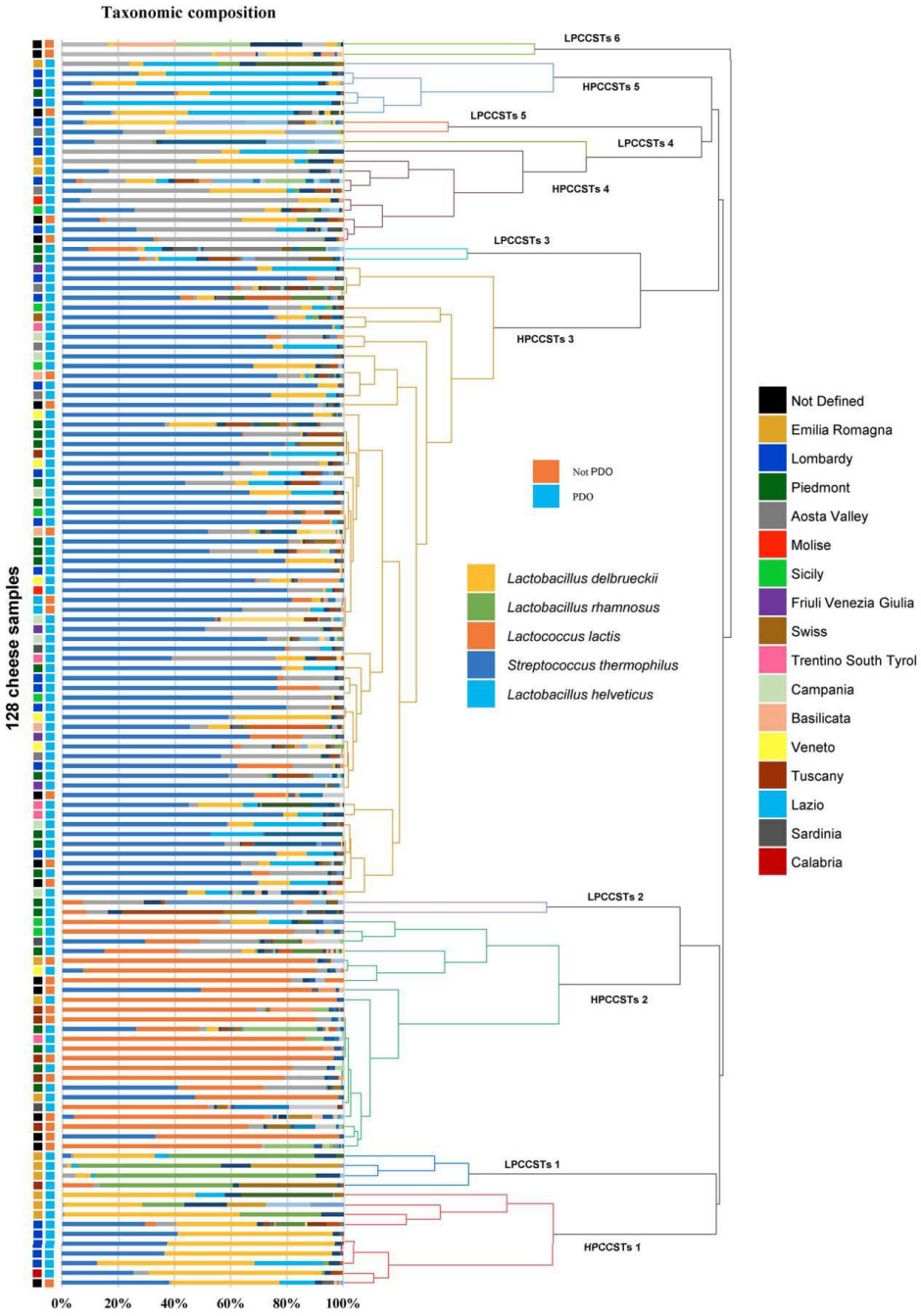
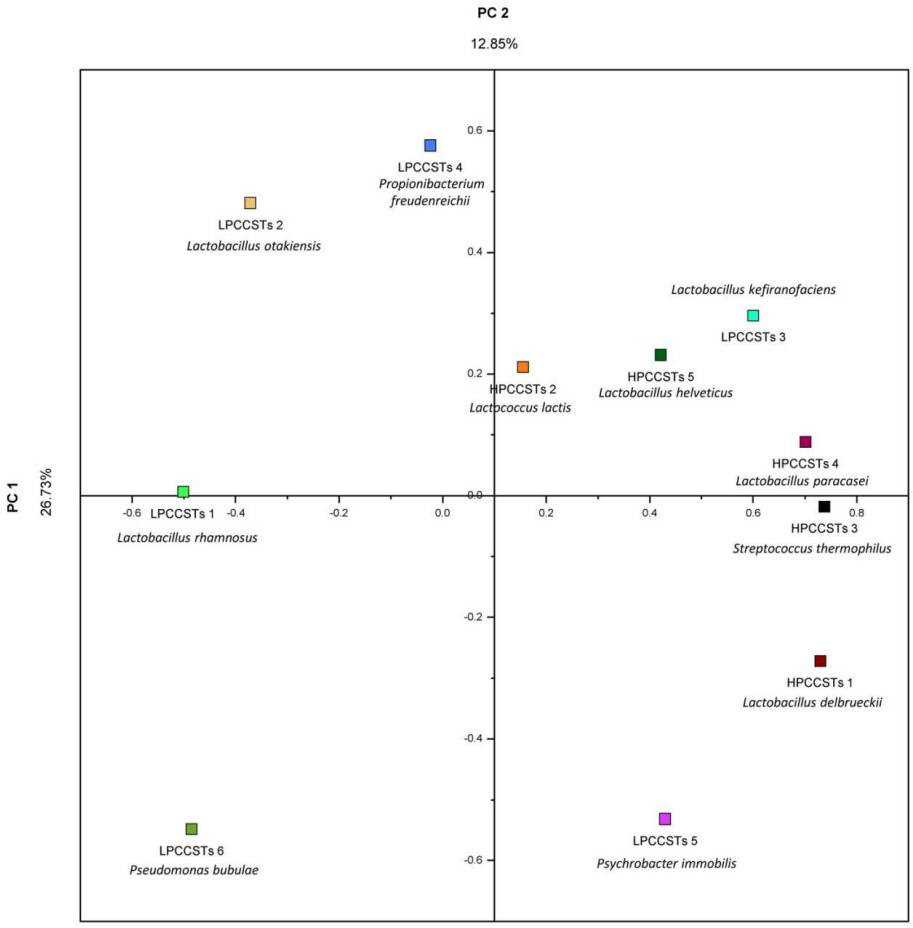


Fig. 2. HCL subdivision of all cheese samples. Graphic representation of HCL subdivision of cheese samples is reported, with branch colored based on HCA cluster. In addition, a stylized

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 bacterium (LAB) (23). 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 and and3;3; Fig. S4) (Data Set S1).



	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

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 and and3;3; Fig. S4) (Data Set S1). In detail, LPCCSTs 5 and 6 represent clusters of contaminants that can be typically found in dairy production (Fig. 2 and and3)3) (Data Set S1) (24, 25). 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 among the most common species exploited as artificial microbial starters in cheese manufacturing (26, 27) (Fig. S5) (Data Set S1). 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 (Fig. S5). 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 ($P < 0.05$) (Data Set S1).

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 microgeographical 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 Text S1).

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 1,746 enzymatic reactions that showed >5% prevalence between the pool of 128 cheese samples analyzed. Because 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 (28) that participate in the establishment of the cheese's organoleptic features and correlate with at least one of the HPCCSTs (26,–31) (Data Set S1) (Fig. 4). The selection of these 48 enzymatic reactions from the correlation pool was

performed manually, exploiting what is reported in the recent literature (32,–35) 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) (Data Set S1). Additional information concerning the selected enzymes and their correlation score with the HPCCSTs are available in the Data Set S1.

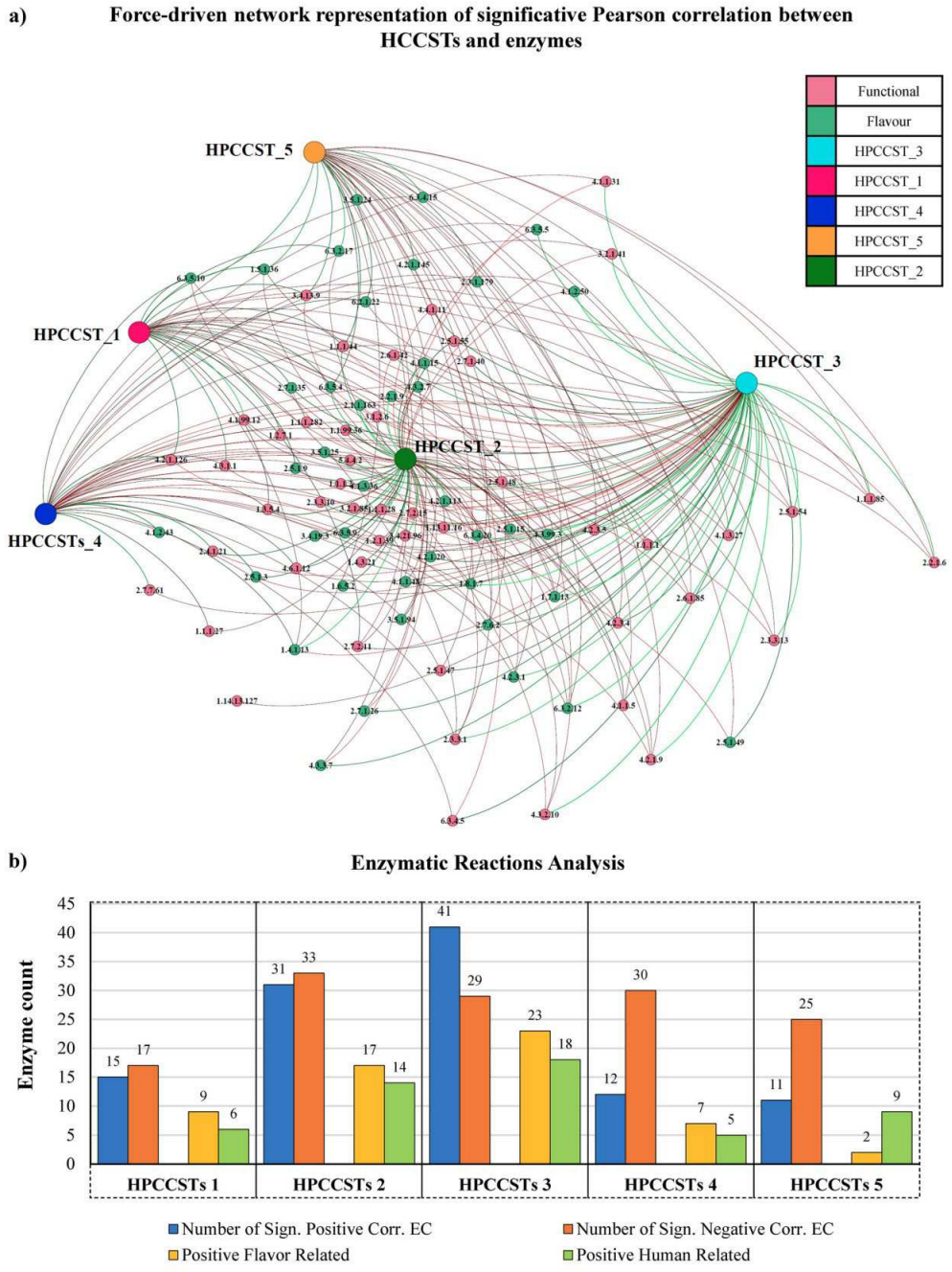


Fig. 4. Human and flavor EC reports. (a) Network representation of correlation analysis based on a significative statistical relationship between the EC – numbers (enzymes) and HPCCSTs. Additionally, nodes were colored in order to separate flavor (green) and human health-related

(pink) enzymes. (b) Barplot 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 (Data Set S1). 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.

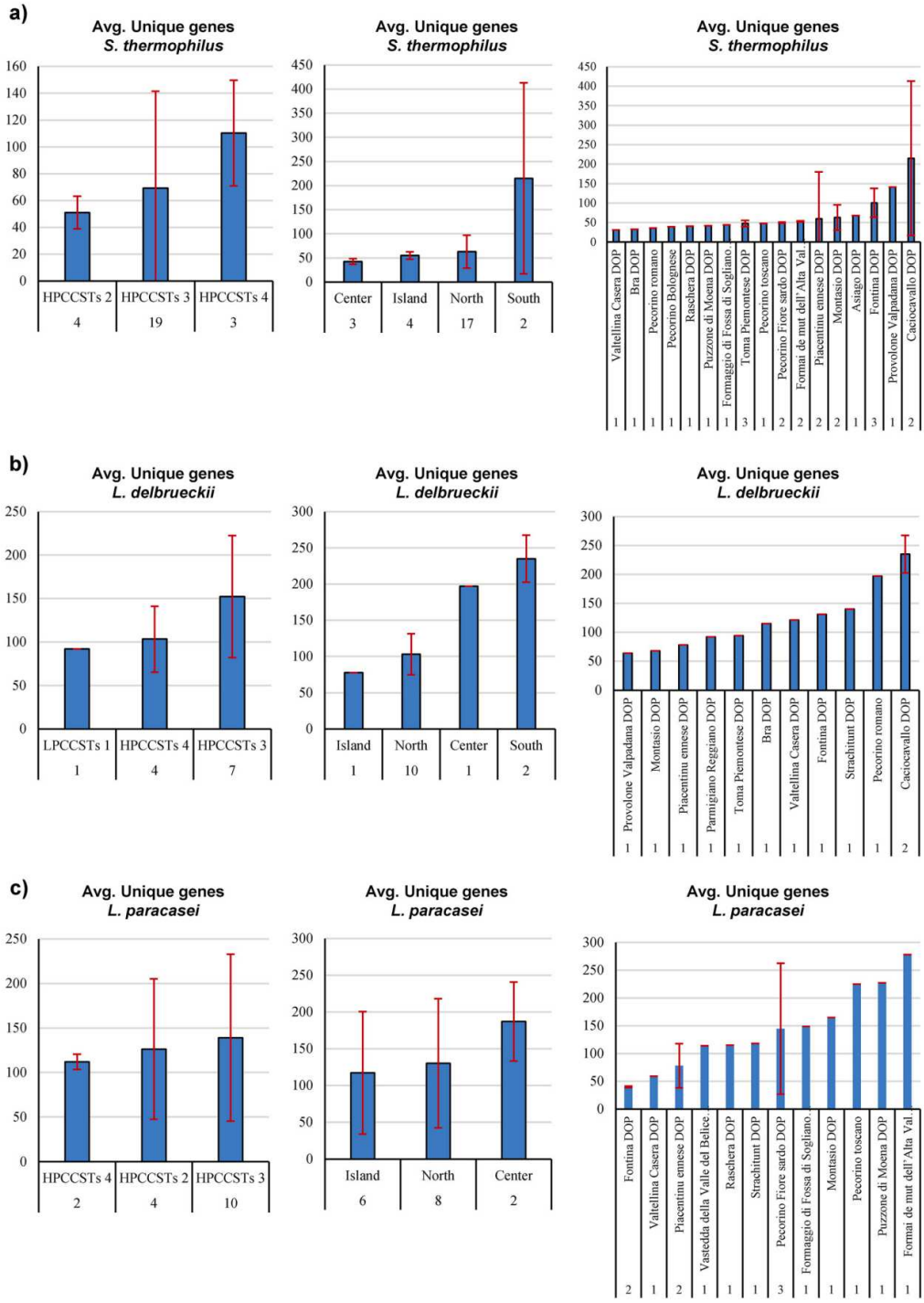

a)
Avg. Unique genes
S. thermophilus
HPCCSTs 2
HPCCSTs 3
HPCCSTs 4
4
19
3
Avg. Unique genes
S. thermophilus
Center
Island
North
South
3
4
17
2
Avg. Unique genes
S. thermophilus
Valtellina Casera DOP
Bra DOP
Pecorino romano
Pecorino Bolognese
Raschera DOP
Puzzone di Moena DOP
Formaggio di Fossa di Sogliano.
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 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

Fig 5. Comparative genomics analysis on unique genes content and metadata subdivision. (a) 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 standard deviation reported when possible. (b) 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 standard deviation reported when possible. (c) 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 standard deviation reported when possible.

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 (Data Set S1) (Fig. 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 (Data Set S1).

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) (Data Set S1). 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 (36). Moreover, lactic acid bacteria (LAB) can also accumulate important secondary metabolites into cheese products, making them a natural supplement of important fermentation by-products (27, 31). 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. Therefore, a subset of 40 enzymatic reactions which showed statistically relevant correlation and that lead to the production of high-interest microbial metabolites (37) was extracted (Data Set S1) (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) (38,–40). 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) (41,–43) and molecules that can participate in the production of GABA (4-aminobutanoate and l-glutamate) (44, 45). 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 (Data Set S1).

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 (Fig. S6). Notably, 71 genomes were selected as they fulfill 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 (Data Set S1) (Fig. S6).

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 (Data Set S1) (Fig. 5). Subsequently, PGAP pipeline (46)

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 (Data Set S1). 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 (standard deviation [SD] of 83.3), 130 (SD of 87.9) and 187 (SD 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 (Fig. S7).

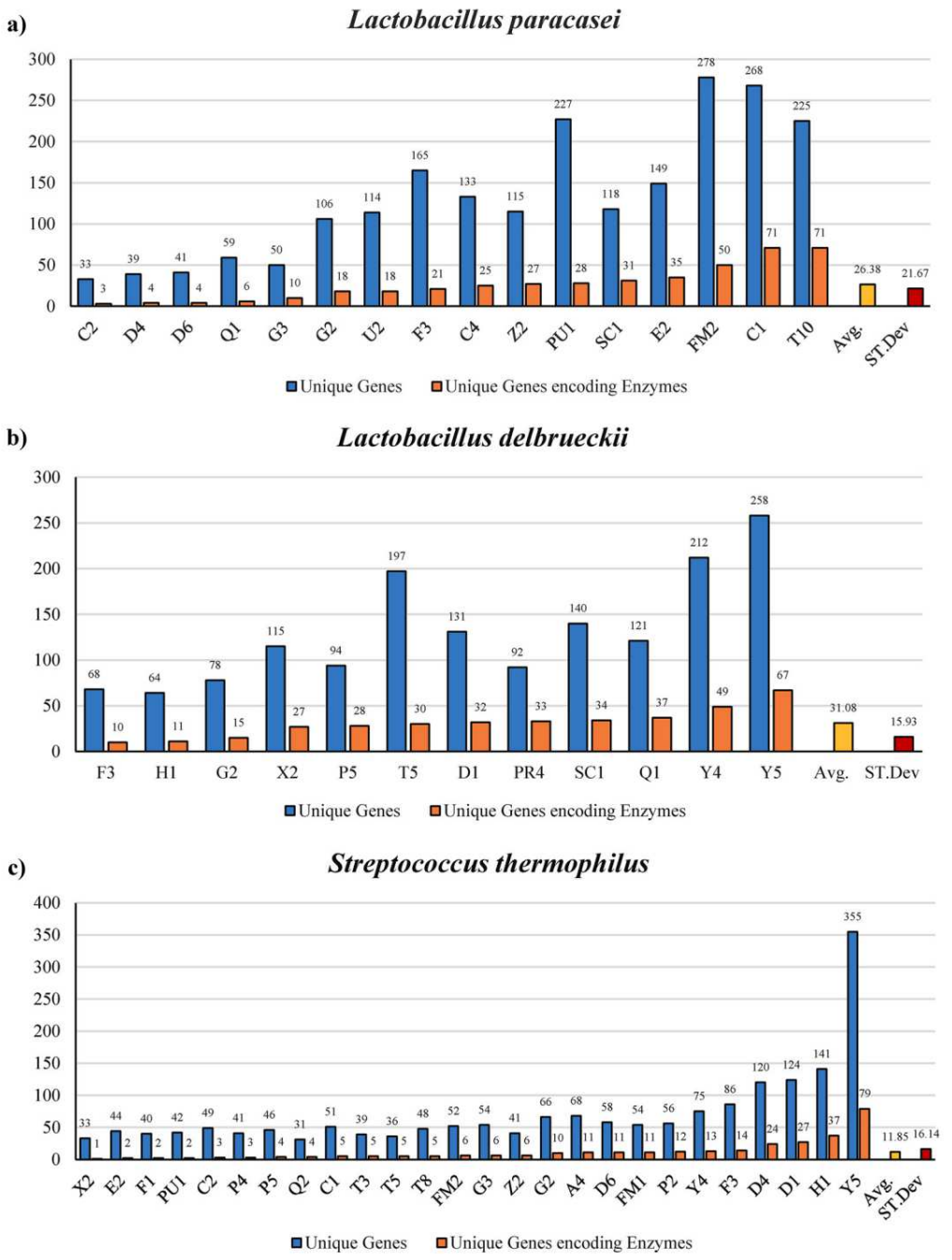


Fig. 6. Unique genes content and enzymatic unique potential. (a) Bar plot showing the unique genes content (blue bar) for each different *L. paracasei* genomes tested, along with the unique

genes encoding enzymes count (orange bar), the average unique genes encoding enzymes count (yellow bar) and the average standard deviation (red bar). (b) 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 standard deviation (red bar). (c) 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 standard deviation (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 with 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. Therefore, 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, ensuring that manufacturing 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 manufacturer-specific 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. 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. All samples 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 to 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, 47). 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 preprocessed 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, 47) 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, each culture replicate 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(36). All data were statistically analyzed with the Attune NxT flow cytometry software.

Statistics and Cluster analysis

HCL analysis was performed on OriginLabPro 2021b (49) 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 (50). 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 (51) software. Three- and two-dimensional PCoA representation of eigenvalues scores was made with OriginLabPro 2021b. PERMANOVA statistical analysis was performed on Rstudio (51) software. One-way ANOVA and independent T-test were performed on SPSS software (52) with 1,000 bootstraps. Pearson bivariate analysis was performed with Rstudios software and represented through a correlation Network made with Gephi software using Force Atlas 2 algorithm (53).

Comparative genomics analysis

Genome quality assessment was performed manually and through the use of checkM (54) software for completeness and contamination score, fastANI (55) software for the Average Nucleotide Identity between strains of the same species and sourmash (56) software for k-mer based genomes comparison. The pangenome and genes orthologous cluster analysis was performed through PGAP (57) software with –identity 0.5 and –coverage 0.8 as set up. DIAMOND (58) 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.

Acknowledgements

We thank GenProbio srl for financial support of the Laboratory of Probiogenomics. Part of this research has been conducted using the high-performance computing (HPC) facility of the University of Parma. This work was financially supported by a postdoc fellowship (Bando 413 Ricerca Finalizzata) to G.A. In addition, F.T. is funded by the Italian Ministry of Health through the Bando 414 Ricerca Finalizzata (Grant Number GR-2018-12365988).

Contributions

F.F. performed bioinformatics and statistical analyses and wrote the manuscript; G.L. collected and sequenced the samples and wrote the manuscript; L.M., G.A.L., C.T., and G.A. managed the metadata and data results; A.V. and R.A. managed the sequencing of samples; F.T. and C.M. supervised the project and edited the manuscript; M.V. supervised the project and designed the study.

Competing interests

The authors declare no competing interests.

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. Cheeses PDO & PGI. https://www.dopitalianfood.com/en/brands-dop-italian-food/cheeses-pdo-pgi.html. Retrieved 2 August 2022.
18. 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.
19. 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.
20. Wolfe BE, Button JE, Santarelli M, Dutton RJ. 2014. Cheese rind communities provide tractable systems for in situ and in vitro studies of microbial diversity. Cell 158:422–433.
21. 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.

22. Enzyme Nomenclature. https://iubmb.qmul.ac.uk/enzyme/. Retrieved 6 April 2022.
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. Smid EJ, Kleerebezem M. 2014. Production of Aroma Compounds in Lactic Fermentations. Annu Rev Food Sci Technol 5:313–326.
25. 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.
26. 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.
27. 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.
28. 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.
29. 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.
30. 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.
31. 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.
32. 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.

33. 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.
34. Tunick MH, Van Hekken DL. 2015. Dairy Products and Health: Recent Insights. J Agric Food Chem 63:9381–9388.
35. 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.
36. 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.
37. 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.
38. RStudio | Open source & professional software for data science teams - RStudio. https://www.rstudio.com/. Retrieved 20 July 2022.
39. Software SPSS - Italia | IBM. https://www.ibm.com/it-it/analytics/spss-statistics-software. Retrieved 25 February 2021.
40. 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.
41. 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.
42. Brown CT, Irber L. 2016. sourmash: a library for MinHash sketching of DNA. J Open Source Softw 1:27.
43. Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J. 2012. PGAP: Pan-genomes analysis pipeline. Bioinformatics 28:416–418.

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

Chapter 7

Designation of optimal reference strains representing the infant gut bifidobacterial species through a comprehensive multi-omics approach.

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

The results of this chapter were published in Environmental Microbiology, 2022 Sep 19; https://doi.org/10.1111/1462-2920.16205.

[§] These authors contributed equally.

[*] These authors contributed equally.

Summary

The genomic era has resulted in the generation of a massive amount of genetic data concerning the genomic diversity of bacterial taxa. As a result, the microbiological community is increasingly looking for ways to define reference bacterial strains to perform experiments that are representative of the entire bacterial species. Despite this, there is currently no established approach allowing a reliable identification of reference strains based on a comprehensive genomic, ecological, and functional context. In the current study, we developed a comprehensive multi-omics approach that will allow the identification of the optimal reference strains using the *Bifidobacterium* genus as test case. Strain tracking analysis based on 1664 shotgun metagenomics datasets of healthy infant faecal samples were employed to identify bifidobacterial strains suitable for *in silico* and *in vitro* analyses. Subsequently, an ad hoc bioinformatic tool was developed to screen local strain collections for the most suitable species-representative strain alternative. The here presented approach was validated using in vitro trials followed by metagenomics and metatranscriptomics analyses. Altogether, these results demonstrated the validity of the proposed model for reference strain selection, thus allowing improved in silico and in vitro investigations both in terms of cross-laboratory reproducibility and relevance of research findings.

Introduction

Historically, the term ‘type strain’ is employed to denote bacterial clonal descendants of the first isolated member of a novel species, also indicated as ‘nomenclatural type’ (Parker et al., 2019). Consequently, type strains were commonly used to perform pairwise comparisons for the assignment of bacterial names, acting as a milestone for taxonomical definition of a microbial taxon (Lapage et al., 1992).

Nevertheless, type strains of most known microbial species may not represent the most suitable candidates to represent their species from an ecological, genomic and functional viewpoint (Kyrpides et al., 2014). For these reasons, the rapid progress in investigating the microbial composition of the human gut microbiota has prompted the need to identify a model strain to be used for experiments to generate results that are representative of the entire bacterial species, thus defining the concept of reference strain (Parker et al., 2019). However, the reference strain definition is somewhat general, indicated as a strain used in comparative studies or any strain derived from a recognized culture collection (Kyrpides et al., 2014; Parker et al., 2019). It is therefore important to identify suitable and relevant microbial reference strains to represent key bacterial species of the human gut microbiota by employing a comprehensive multi-omics approach.

Among the hundreds of different bacterial species inhabiting the human intestine, members of the genus *Bifidobacterium* have received substantial research consideration in recent decades. Bifidobacteria are among the first colonizers of the infant gut, persisting at high abundance at least until the weaning phase, while certain species/strains of this genus are claimed to exert multiple beneficial

effects on host health (Alessandri et al., 2019, 2021; Bottacini et al., 2017; Hidalgo-Cantabrana et al., 2017; Milani et al., 2016; Rivière et al., 2016; Turroni, Milani, Duranti, Mahony, et al., 2018). In this context, various studies have investigated the bifidobacterial community present in the infant intestine through metagenomic analyses of faecal samples, revealing that *Bifidobacterium bifidum*, *Bifidobacterium breve*, *Bifidobacterium longum* subsp. *infantis* and *B. longum* subsp. *longum* represent the most prevalent and abundant bifidobacterial species of the 'infant-like' gut microbiota (Arboleya et al., 2016; Duranti et al., 2017; Milani et al., 2015; Tarracchini et al., 2022; Turroni et al., 2021; Turroni, Milani, Duranti, Ferrario, et al., 2018). Furthermore, scientific interest in bifidobacteria has led to extensive isolation efforts that now allow access to a large number of strains and their corresponding genomic sequences (Saturio et al., 2021; Turroni et al., 2022), rendering this genus a suitable example to perform phylogenomic studies.

There is currently no established approach to appropriately assign reference strains on the basis of genetic/genomic evidence, which should more accurately represent the typical biological features that characterize each bifidobacterial species. Notably, in most cases, selection of bifidobacterial strains for in vitro and in vivo studies is not based on current physiological, ecological and genetic knowledge, possibly causing inaccurate or confusing insights (Arboleya et al., 2016; Duranti et al., 2017; Milani et al., 2015; Turroni, Milani, Duranti, Ferrario, et al., 2018). In fact, several bifidobacterial comparative genomic analyses have reported remarkable intra-species genetic variability, principally due to the presence of a high average number of truly unique genes, that is, genetic sequences encoded by a single genome, within a single bifidobacterial species.

In addition, these highly variable genetic traits are responsible for metabolic and physiological differences among strains of a given species, which may complicate experimental consistency of traits that could be applied to the designation of valuable reference strains for each bifidobacterial specie (Bottacini et al., 2018; Lugli et al., 2020; Lugli, Duranti, Albert, et al., 2019; Lugli, Mancino, Milani, et al., 2019; Tarracchini et al., 2021).

To address this issue, we assessed the abundance of publicly available strains belonging to bifidobacterial species harboured by the infant gut through a very detailed strain-genomic reconstruction of the infant gut microbiomes. Strain profiling data were then coupled with genomic and comparative genome analyses to define the most representative strains of each bifidobacterial species in the collected infant gut microbiomes. We then developed a bioinformatic tool, here referred to as RefBifSelector, which performs a phylogenetically based rapid screening of the local bifidobacterial strain collections, allowing the identification of the 'most suitable and locally available bifidobacterial strain(s)' based on the closest genomic relative to the newly assigned reference bifidobacterial strains. The validity of the identified novel representative bifidobacterial model strains in terms of microbe–microbe and microbe–host interactions was investigated by means of metagenomic and metatranscriptomic methods. The latter approaches were applied using in vitro models simulating real-life settings involving (i) co-occurrence with complex bacterial communities encompassing other members of the infant gut microbiota and (ii) molecular interactions with human intestinal cells.

Experimental Procedures

Dataset selection

All publicly available datasets corresponding to infant faecal samples sequenced through a shotgun metagenomics approach were selected and downloaded from the NCBI SRA repository using NCBI SRA Toolkit 2.11.0 faster-dump (Leinonen et al., 2011). Specifically, only datasets processed through Illumina sequencing technology were retained to achieve high quality and coverage data. Additionally, only shotgun metagenomic datasets belonging to healthy infants and full-term infants with age ranging from a few days to 3 years of life and not having undergone drug, probiotic and/or prebiotic treatments were included in this study. Conversely, no exclusion criteria based on diet, mode of delivery or geographical origin were applied for the selection of the final infant cohort.

Shotgun metagenomics dataset analysis

To analyse only high-quality SRA samples, each SRA was subjected to a filtering step to remove 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).

Collected filtered reads were then taxonomically classified through the METAnnotatorX2 pipeline (Milani et al., 2021), using the up-to-date RefSeq (genome) database retrieved from the NCBI. Metagenome-assembled genomes (MAGs) were reconstructed through the METAnnotatorX2 pipeline (Milani et al., 2021) and (meta)SPAdes software (Bankevich et al., 2012). Reconstructed contig >5 kbp were retained and taxonomically classified in the same manner as for filtered reads.

Bifidobacterial strain selection.

To create species-specific databases comprising bifidobacterial species typical of the infant gut microbiota, filtered reads from online shotgun metagenomics datasets were subjected to whole metagenome assembly using SPAdes v3.14 (Bankevich *et al.*, 2012) with default parameters and the metagenomic flag option (-meta) together with minimum k-mer sizes of 21, 33, 55, and 77, as previously described (Lugli, Duranti, Milani, *et al.*, 2019; Lugli, Milani, *et al.*, 2019; Lugli *et al.*, 2021). For short reads, reconstructed contig sequences were taxonomically classified based on their sequence identity using megablast against the same RefSeq database (Agarwala *et al.*, 2018). Furthermore, in addition to bifidobacterial genomes reconstructed from shotgun metagenomic datasets, publicly available genomes belonging to *Bifidobacterium* species typical of the infant gut microbiota were selected from the NCBI genome list. In detail, only genome sequences with a genome coverage higher than 30-fold and containing less than 100 contigs were considered.

StrainGST-based creation of bifidobacterial species-specific databases.

The genomes of the selected bifidobacterial strains, both those reconstructed from online datasets and those publicly available and new isolates, were used as input in the StrainGE employing StrainGST analysis (van Dijk et al., 2021) with standard parameters to identify possible redundant genomes among those selected for database creation. In this context, to remove genomes with sequence identity >99% from final databases, straingGST function 'straingst kmersim' first calculated Jaccard similarity between k-mer profiles of each genome pair. K-mer profiles evaluate the distribution and prevalence of short sequences, that is, k-

mers, across the genomic sequences. This heuristic approach, extensively used in modern genomic and metagenomic analyses, provides a marked reduction in computing time and computational resources needed (Bernard et al., 2018). Subsequently, the function 'straingst cluster' was used to remove all genomes which shared 99% of the k-mers with another genome already included in the database. Additionally, straingGST was exploited to generate clusters of similar genomes with Jaccard similarity between two k-mer sets higher than 0.90, selecting only the optimal one inside each cluster. Hierarchical clustering (HCL) analysis was performed through further neighbourhood, Pearson correlation and maximum distance. RefBif-IS was then selected using a specific index, called AxP, defined as [the average ANI value of genomes constituting the same HCL] * [prevalence score of the strain in the IGMC] * [100].

Ethical statement

The study protocol was approved by the Ethical Committee of the "Azienda Unità Sanitaria Locale di Reggio Emilia – IRCCS" in Reggio Emilia, Italy. A signed informed consent was obtained from the legally authorized representative of the infant enrolled in this study.

Experimental set up for infant gut microbiota stabilization

A faecal sample from a 3-year-old healthy infant who had not been treated with antibiotics, prebiotics, or probiotics in the previous 3 months was collected immediately after defecation and transported under anaerobic conditions to the laboratory where it was immediately processed. Specifically, upon receipt, the faecal sample was transferred into an anaerobic chamber and immobilized in 1–

2-mm-diameter gel beads composed of 2.5% (w/v) gellan gum, 0.25% (w/v) xanthan gum, and 0.2% (w/v) sodium citrate, as previously described (Cinquin et al., 2004; Le Blay et al., 2010; Pham et al., 2019).

The fermentation medium was based on the composition designed to mimic the infant gut environment as previously described (Zihler Berner *et al.*, 2013). Specifically, the medium contained the following components: 5 g L^{-1} starch, 2 g L^{-1} pectin, 1 g L^{-1} guar gum, 4 g L^{-1} mucin from porcine stomach, 2 g L^{-1} xylan, 2 g L^{-1} arabinogalactan, 1 g L^{-1} inulin, 3 g L^{-1} casein, 5 g L^{-1} peptone water, 5 g L^{-1} tryptone, 0.4 g L^{-1} bile salts, 0.005 g L^{-1} $FeSO_4$, 4.5 g L^{-1} NaCl, 0.5 g L^{-1} KH_2PO_4, 0.61 g L^{-1} $MgSO_4$, 0.1 g L^{-1} $CaCl_2$, 1.5 g L^{-1} $NaHCO_3$, 0.8 g L^{-1} cysteine, 0.2 g L^{-1} $MnCl_2$, 0.05 g L^{-1} hemin, and 1 ml Tween 80 (Macfarlane et al.; Zihler Berner et al.). Final pH was adjusted to 6.8, while, after autoclaving, a 0.2 μm filter-sterilized vitamin solution (1 mL L^{-1}) was added to the medium (Pham *et al.*, 2019). The nutritive medium was freshly prepared daily, autoclaved, and stored at 4°C under stirring until use.

The fermentation setup consisted of a bioreactor (Solaris Biotech Solutions, Italy) with a working volume of 400 ml inoculated with 40 ml (10% v/v) faecal beads. Fermentation started in a batch mode by aseptically replacing spent medium with fresh medium every 12 h for 3 days, after which the fermentation was switched to a continuous mode by feeding the bioreactor with fresh medium at flow rates of 66 ml h−1 to obtain a mean retention time of 6 h, as previously described (Doo et al., 2017). Stabilization of the infant gut microbiota was performed under anaerobic conditions, while temperature was set at 37°C, stirring speed at 180 rpm, and pH was maintained automatically at 6.8 by adding 2.5 M NaOH. The fermentation process was carried out for a total of 10 days.

Bifidobacterial cultivation in a gut simulated environment

To evaluate the in vitro ability of bifidobacterial strains selected through the RefBifSelector tool among our repository to grow in an intestinal environment, *B. adolescentis* 713B, *B. bifidum* PRL2010, *B. breve* 1895B, *B. dentium* 181B, *B. longum* subsp. *longum* 39B and *B. pseudocatenulatum* 1896B, as well as the type strains and the identified sub-optimal strains for each selected bifidobacterial species were in batch cultivated in a complex medium mimicking the infant intestine (Macfarlane et al., 1998) together with the abovementioned stabilized gut microbial community of a 3-year-old infant.

Specifically, bifidobacterial strains were revitalized from glycerol-based stock in MRS broth medium supplemented with 0.05% (w/v) l-cysteine hydrochloride at 37°C under anaerobic conditions. Subsequently, bifidobacterial strains together with the 10-day stabilized infant gut microbial community were singularly inoculated in 45 ml of culture medium mimicking the infant intestinal environment (Macfarlane et al., 1998; Zihler Berner et al., 2013) to obtain a final inoculum of 105 cells ml−1 of a specific bifidobacterial strain and 107 cells ml−1 of the stabilized infant gut microbial community. For each experiment, an aliquot of culture was collected at three different time points, that is, 6, 12, and 24 h after the inoculum and conserved at −20°C until they were processed for DNA extraction and flow cytometry-based bacterial total count. Batch cultures were incubated in an anaerobic chamber (2.99% H2, 17.01% CO2 and 80% N2) (Concept Ruskinn) at 37°C.

DNA extraction and shallow shotgun sequencing

Each aliquot of every single obtained batch fermentation, together with the control sample, that is, the 10-day stabilized infant gut microbial community were subjected to DNA extraction using the QIAmp DNA stool mini kit following the manufacturer's instructions (Qiagen, Germany). The extracted DNA was prepared using the Illumina Nextera XT DNA Library Preparation Kit and following the Illumina NexteraXT protocol. Specifically, DNA samples were enzymatically fragmented, barcoded, and purified using magnetic beads. Subsequently, samples were quantified using a fluorometric Qubit quantification system (Life Technologies, USA), then loaded on a 2200 Tape Station Instrument (Agilent Technologies, USA) and normalized to 4 nM. Paired-end sequencing was performed using an Illumina MiSeq sequencer with MiSeq Reagent Kit v3 (Illumina Inc., San Diego, USA). Taxonomic reconstruction of shallow shotgun data was performed as described above for shotgun metagenomic data analysis.

Evaluation of bacterial cell density by flow cytometry

For total cell counts, each culture replicate was 100,000 times diluted in physiological solution (phosphate-buffered 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 min before measurement. All enumeration experiments were carried out through 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, that

is, 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. In addition, 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. All data were statistically analysed with the Attune NxT flow cytometry software.

Human cell line culture

Human colorectal carcinoma-derived Caco-2 cells (purchased from ATCC) and HT29-MTX (kindly provided by Prof. Antonietta Baldi, University of Milan), i.e., a human colon carcinoma-derived, mucin-secreting goblet cell line, were cultured in Minimum Essential Medium (MEM) and Dulbecco's Modified Eagle's medium (DMEM) with high glucose (4.5 g/L) and 10 mM of sodium pyruvate, respectively, as previously described (Doo *et al.*, 2017). In addition, both media were supplemented with 10 % Fetal Bovine Serum (FBS), 2 mM glutamine, 100 g/mL streptomycin, and 100 U/mL penicillin. Cultures were maintained at 37°C in a humidified atmosphere of 5 % CO_2 in air in 10-cm dishes and passaged three times a week. Subsequently, a mixed suspension of Caco-2 and HT29-MTX cells (7:3) was seeded in DMEM + FBS at a density of $\approx 10^5$ cells/cm^2 into cell culture inserts with membrane filters (pore size 0.4 μm) for Falcon 24-well-multitrays (Becton, Dickinson & Company, Franklin Lakes, NJ, USA), and cultured for 21 days with a medium replacement every three days until a tight monolayer was formed (TEER > 600 $\Omega \bullet cm^2$).

Human cell-monolayers and bifidobacteria

After 21 days from seeding, the culture medium of the 24-well plates was replaced with fresh antibiotic-free DMEM. Subsequently, bifidobacterial cells (final concentration 108 cells ml^{-1}) were added to Caco-2/HT29-MTX cell monolayers, as previously described (Turroni et al., 2014). The 24-well plates were then incubated in 5% CO2 at 37°C for 4 h. After this period of incubation, bacterial cells were recovered in RNA later and stored at −80°C until processing. For these experiments, each bifidobacterial strain, that is, B. bifidum PRL2010, B. breve 1895B, B. longum subsp. longum 39B, and B. pseudocatenulatum 1896B, together with the Type Strains and the sub-optimal strains of the same bifidobacterial species, were grown in MRS broth in anaerobic chamber at 37°C. Once they reached the exponential phase of growth ($0.6 < OD600nm < 0.8$), cells were enumerated through Thoma cell counting chamber (Herka), possibly diluted to reach a final concentration of 108 cells ml^{-1}, washed in PBS, resuspended in DMEM without antibiotics, and seeded on Caco2/HT29-MTX cell monolayers. Furthermore, for each bifidobacterial strain, a control sample, that is, the bifidobacterial strain resuspended in DMEM and maintained under the same incubation conditions of the 24-well plates without any contact with human cell lines, was obtained.

RNA extraction

Total RNA of each considered condition was isolated using a previously described method (Milani et al., 2020; Turroni et al., 2016). Briefly, bifidobacterial cell pellets were resuspended in 1 ml of QIAzol lysis reagent (Qiagen) in a sterile tube containing glass beads (Merk, Germany). Cells were

lysed by alternating 2 min of stirring the mix on a Precellys 24 homogenizer (Bertin instruments, France) with 2 min of static cooling in ice. These steps were repeated three times. The lysed cells were centrifuged at 12,000 rpm for 15 min, and the upper phase was recovered. The RNA samples were then purified using the RNeasy minikit (Qiagen) following the manufacturer's instruction. RNA concentration and purity were evaluated using a spectrophotometer (Eppendorf, Germany).

RNA sequencing analysis

For RNA sequencing, total RNA (from 100 ng to 1 μg) was treated to remove rRNA by using the QIAseq FastSelect—5S/16S/23S following the manufacturer's instructions (Qiagen). The yield of rRNA depletion was checked by using a 2200 TapeStation (Agilent Technologies, USA). Then, a whole transcriptome library was constructed using the TruSeq Stranded mRNA Sample preparation kit (Illumina). Samples were loaded into a NextSeq high-output v2 kit (150 cycles) (Illumina) as indicated by the technical support guide. The obtained reads were filtered to remove low-quality reads (minimum mean quality 20 and minimum length 150 bp) as well as any remaining ribosomal loci using the METAnnotator X2 pipeline (Milani et al., 2021) Subsequently, the retained reads were aligned to the specific reference genome through Bowtie2 software (Langdon, 2015). Analysis of the RPKM values was performed through Artemis using the formula RPKM (Reads Per Kilobase Million) = numReads/(geneLength/1000 * totalNumReads/1000000) (Mortazavi et al., 2008).

Results and Discussion

Ecological and phylogenomic-driven identification of optimal reference model strains.

In order to analyse the ecological distribution of the most representative bifidobacterial strains of the infant gut microbiota, we established an infant faecal sample database called Infant Gut Microbiota Collection (IGMC), consisting of metagenomic data sets of faecal samples from individually selected healthy infants, ranging in age from a few days to 3 years, sequenced through Illumina-based technology in order to avoid any bias related to different sequencing platforms (Figure 1) (Table S1) Table 1. Based on these criteria, a total of 1664 datasets covering 17 different cohorts from various geographical areas were selected and analysed through METAnnotatorX2 (Milani et al., 2021) (Table 1).

TABLE 1

SRA metadata summary report

Bioproject	SRA count	% of IGMC (%)	Age (average days)	Country	Count	Prevalence in IGMC (%)
PRJEB12669	14	0.84	544	Europe	593	35.64
PRJEB24771	233	14.00	183	USA	554	33.29
PRJEB32135	65	3.91	252	Malawian	233	14.00
PRJEB6456	199	11.96	239	New Zealand	125	7.51
PRJNA287207	2	0.12	90	Canada	65	3.91
PRJNA290380	394	23.68	545	South Africa	56	3.37
PRJNA322188	32	1.92	44	Italy	24	1.44
PRJNA339914	5	0.30	90	China	14	0.84
PRJNA345144	125	7.51	362			
PRJNA352475	14	0.84	120			
PRJNA422569	3	0.18	730			
PRJNA473126	56	3.37	235			
PRJNA475246	60	3.61%	90			
PRJNA524703	306	18.39	156			
PRJNA542703	14	0.8	45			
PRJNA549787	56	3.37	113			
PRJNA557731	86	5.17	30			

Abbreviation: IGMC, Infant Gut Microbiota Collection.

Additional information regarding the IGMC can be found in the Supplementary Text S1.

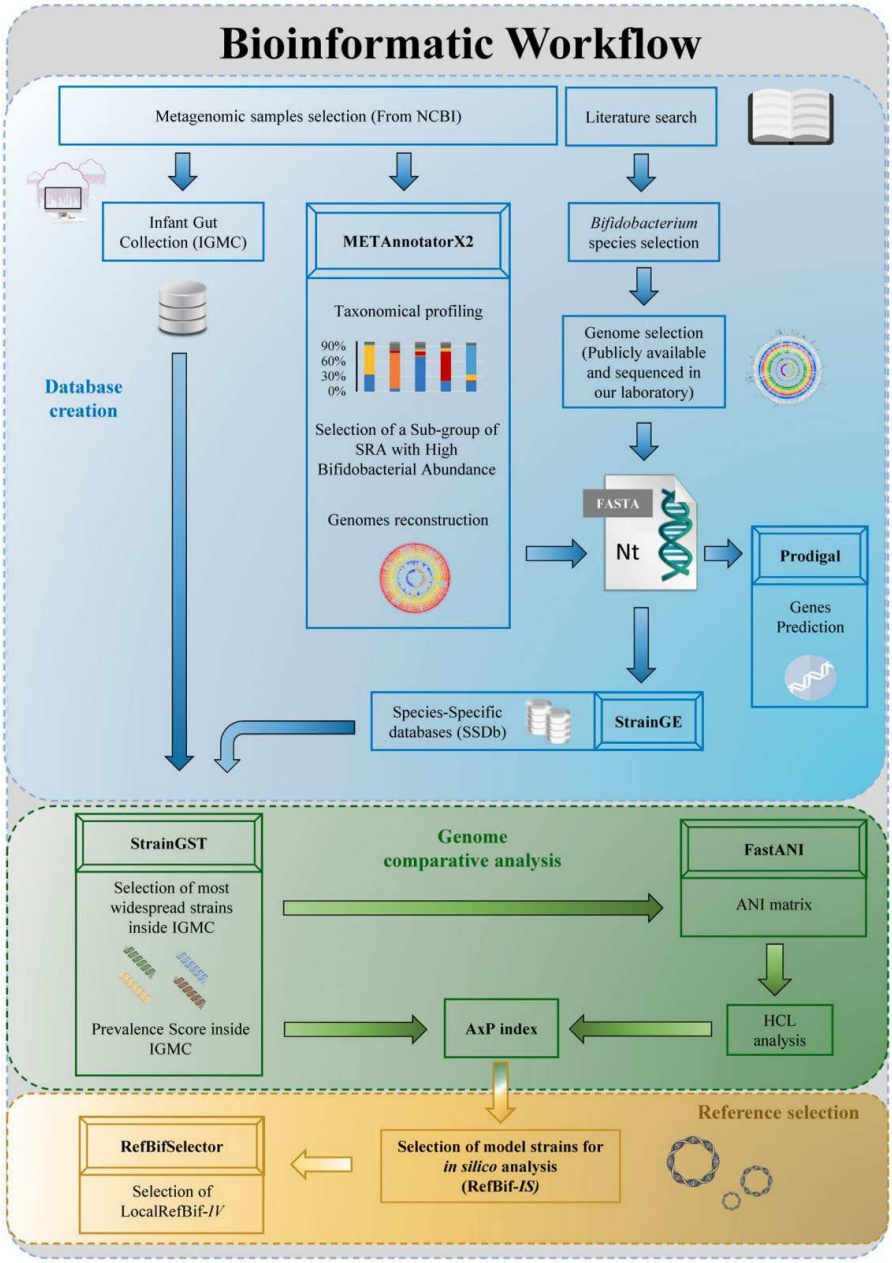


Figure 1. Bioinformatics workflow. This figure depicts a schematic workflow concerning the main bioinformatics steps performed. The workflow is divided into three main blocks, outlining steps for data recovery, their analysis and the creation of databases (blue block), the analyses on selected genomes (green block), and the final selection of RefBif-IS and LocalRefBif-IV (yellow block).

The IGMC dataset was scrutinized to define the most abundant and prevalent bifidobacterial species, representing the bifidobacterial 'core' infant gut microbiota. Specifically, this bifidobacterial core community was defined by selecting those bifidobacterial taxa showing a prevalence of >10% and an average abundance of >0.5%, so as to consider only those species playing a relevant role in the infant gut microbiota (Turroni et al., 2021). In this context, *B. adolescentis*, *B. bifidum*, *B. breve*, *B. catenulatum*, *B. longum*, *B. dentium* and *B. pseudocatenulatum* were shown to be the most representative Bifidobacterium species across the IGMC, confirming previously published data (Arboleya et al., 2016; Duranti et al., 2017; Laursen et al., 2021; Milani et al., 2015; Turroni, Milani, Duranti, Ferrario, et al., 2018) (Figure 1) (Table S2) (Details are provided in the Supplementary Text S1).

Subsequently, datasets showing a *Bifidobacterium* genus average relative abundance >10% were selected to reconstruct the genomes of bifidobacterial strains corresponding to the above-identified most prevalent species. The threshold of 10% average relative abundance in bifidobacterial composition was selected to obtain enough genetic material (i.e. reads) to reconstruct (near complete) genomes. This procedure allowed the reconstruction of 239 bifidobacterial metagenomes assembled genomes (MAGs), including 105, 47, 30, 19, 16, 13, and nine chromosomal sequences belonging to the *B. longum*, *B. bifidum*, *B. breve*, *B. pseudocatenulatum*, *B. dentium*, *B. adolescentis*, and B. catenulatum species, respectively (Table S3). These MAGs comprise the dominant strains of a bacterial species along with fragments of sequences derived from other strains belonging to the same bacterial species if present at relatively high abundance.

So, MAGs are microbial reconstructed genomes representative of the bacterial species encompassing the microbiomes of the samples assayed.

These reconstructed MAGs, combined with 965 publicly available bifidobacterial genomes and 93 bifidobacterial strains that had previously been isolated and sequenced within the context of the current study, were then used to generate Species-Specific Databases (SSDbs), in total encompassing 1297 bifidobacterial chromosomes of the abovementioned key microbial players of the infant gut microbiota (Figure 1) (Table S3). This database, which covers the highest genomic diversity available for each bifidobacterial species, was built to evaluate the prevalence of bifidobacterial strains within the previously created IGMC (Tables S3 and S4).

To evaluate the distribution of bifidobacterial strains among the IGMC datasets, a StrainGST-based profiling analysis was conducted, resulting in the identification of 209, 76, 70, 48, 47, 21 and 19 strains belonging to *B. longum*, *B. bifidum*, *B. breve*, *B. adolescentis*, *B. pseudocatenulatum*, *B. dentium* and *B. catenulatum* species, respectively, for a total of 490 bifidobacterial genomes showing a prevalence >0.1% across the IGMC datasets (Figure 1) (Table S5). These 490 bifidobacterial genomes were subjected to ANI analysis to evaluate genome similarity, and the obtained ANI data were subsequently used to perform a hierarchical clustering analysis (HCA) (Liang et al., 2018; OriginLab, 2020), resulting in a series of species-specific HCL trees. (Supplementary excel File S1) (Details are provided in the Supplementary Text S1).

Finally, strain prevalence and ANI data were integrated into a specific index score, that is, the Average × Prevalence index (A × P index), as described in the Experimental Procedures. This score led to the selection of eight reference strains

with the highest A × P scores for each species-specific HCL tree, therefore considered to represent the optimal Reference Bifidobacterial strains for in silico analyses (RefBif-IS) (Supplementary excel File S1) (Figures 1 and 2) (supplementary text S1). Therefore, these eight reference bacterial strains can be considered the most ecologically and genetically representative bifidobacterial strains of the infant gut microbiota (Table 2).

TABLE 2

RefBif-*IS* selected as optimal reference strains

Species	Strain	GCA
Bifidobacterium adolescentis	LMG 10734	GCA_002107995.1
Bifidobacterium bifidum	1001283B150225_161107_H11	GCA_015549125.1
Bifidobacterium breve	GED8481	GCA_001546235.1
Bifidobacterium catenulatum	JCM 1194	GCA_001025195.1
Bifidobacterium dentium	ATCC 27679	GCA_000146775.1
Bifidobacterium longum subsp. *longum*	1-5B	GCA_000730105.1
Bifidobacterium longum subsp. *infantis*	*B.longum*_ssp_infantis_5	GCA_902167565.1
Bifidobacterium pseudocatenulatum	JCM 1200	GCA_001025215.1

Intriguingly, with the exception of *B. pseudocatenulatum* JCM 1200T, all other former Type Strains (or type strain cluster representing from StrainGE analysis) *B. longum* subsp. *longum* ATCC 15707T, *B. longum* subsp. *infantis* ATCC 15697T, *B. dentium* DSM 20436T, *B. catenulatum* DSM 16992T, *B. breve* ATCC 15700T, *B. bifidum* JCM 1255T and *B. adolescentis* ATCC 15703T show a much lower A x P value than the RefBif-IS (Figure 2).

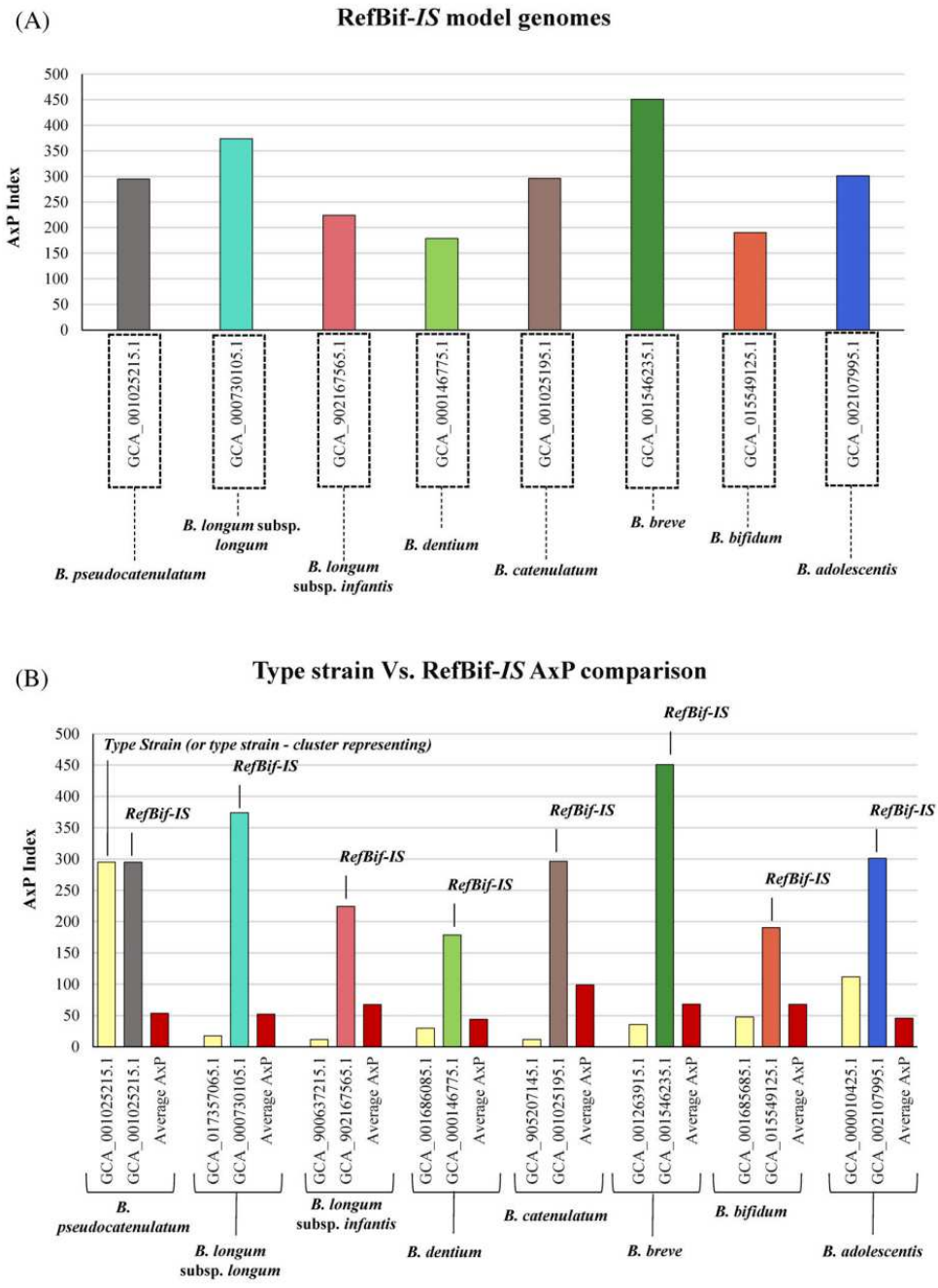


Figure 2. Model strain for in silico and in vitro analyses. Panel (A) reports, in the x axis, the optimal reference strains for in silico analyses identified for each bifidobacterial species typical of the infant gut microbiota, while the y axis shows the AxP index associated with each RefBifIS.

Panel (B) reports a comparison between the type strain versus RefBif-IS and their AxP values, with the average AxP of the total HCL as reference.

This result suggests that these bifidobacterial type strains do not accurately represent the genetic content or ecological distribution of each bifidobacterial species, as also recently suggested (Figure 2) (Lapage et al., 1992). Although this approach allocates equal weight to all genes, this was done to normalize the results, considering the prevalence as genetic adaptation favourable to the environment, thus assuming the weight of all genes simultaneously. In conclusion, these A × P values emphasize that the suitability of the former physiologically defined type strains to act as model organisms for each species should mainly be based on the genetic make-up of microbial species, and thus on their genomic variability (Supplementary excel File S1) (Figure 2).

RefBifSelector: a tool for biobank screening.

While RefBif-IS represents a valuable genetic-driven approach to identify suitable reference strains for each infant bifidobacterial species, it may be that the indicated strains are not publicly available and/or easy to retrieve/obtain. We therefore considered the possibility to expand the selection also to those strains that are available in non-public repositories (e.g. local microbial collections) in order to identify phylogenomically related alternatives that labs may use as alternative reference models. For this reason, we developed a tool named RefBifSelector, which allows a rapid screening of local biobanks for bifidobacterial strains that are closely related to the optimized reference strains defined by RefBif-IS (https://probiogenomics.unipr.it/cmu/) (Figure S1). This tool requires the bacterial genome sequences in fasta format to be screened as

input (Query) and provides the best alternative to the reference RefBif-IS genomes as output. This evaluation is performed through ANI score in conjunction with the use of the average percentage of positive scoring matches (PPOS). While ANI analysis investigates the nucleotide identity between genome pairs, PPOS is a score retrieved from Blastp analysis that uses the following formula to compare the translated amino acid sequences: [(number of identical matches) + (number of similar matches)]/(alignment length). Therefore, a final score equal to the value of ANI * Average_PPOS was obtained, considering a minimum score threshold equal to 9600 (corresponding to an ANI and Average_PPOS of 98). Strains below this minimum threshold cannot be assessed adequately similar to RefBif-IS and cannot be considered suitable choices.

In order to validate this bioinformatic pipeline, genomes of bifidobacterial strains belonging to our local biobank were submitted as input to the RefBifSelector tool. Specifically, among the 34 screened strains, *B. bifidum* PRL2010, B. longum subsp. longum 39B, B. breve 1895B and *B. pseudocatenulatum* 1896B displayed the best choice in terms of ANI and Average PPOS value with respect to the RefBif-IS references, thus representing the LocalRefBif-IV (in vitro), while *B. catenulatum*, *B. adolescentis*, *B. dentium* and *B. longum* subsp. *infantis* were excluded from the analysis since no strains of human origin were available in the bifidobacterial strains repository that we employed (Table S6).

***In vitro* validation of the identified LocalRefBif-*IV*.**

Optimal bacterial models should allow the efficient investigation of microbe–microbe and host–microbe interactions, including those occurring in complex microbial populations.

In this context, to further corroborate the biological suitability of the strains identified through the RefBifSelector tool, we assayed the ability of the identified LocalRefBif-IV and the other members of the infant gut microbiota using an in vitro model. Therefore, the four bifidobacterial strains identified as LocalRefBif-IV among our local bifidobacterial strain collection were cultivated in a simulated infant intestinal environment. Additionally, the respective former bifidobacterial type strains for *B. pseudocatenulatum*, *B. longum, B. breve* and *B. bifidum*, that is, LMG10505, LMG13197, LMG13208 and LMG11041 and four strains of our local biobank identified as the most genetically dissimilar to the RefBif-IS for *B. pseudocatenulatum*, *B. longum*, *B. breve* and *B. bifidum*, that is, 1052B, 209B, 31L and 324B, were cultivated to evaluate and compare their growth performances with that of the four bifidobacterial strains predicted as LocalRefBif-IV, that is, 1896B, 39B, 1895B and PRL2010. Specifically, each selected strain was inoculated in batch culture systems using a complex medium mimicking the infant intestinal environment (Macfarlane et al., 1998; Zihler Berner et al., 2013) together with a healthy 3-year-old infant-derived gut microbial community previously stabilized through a continuous fermentation model. Cultivation samples taken at 6, 12, and 24 h were collected for each strain, together with a single control sample, that is, the inoculated stabilized infant gut microbial community. Subsequently, the microbial DNA extracted from each condition was subjected to shallow shotgun sequencing (Hillmann et al., 2018;

Milani et al., 2021), generating a total of 919,824 reads with an average of 48,411 reads per sample, reduced to a total of 695,150 reads and an average of 36,586 reads per sample after quality-filtering (Table S7).

Taxonomic profiling at the species-level highlighted the presence of traces of B. longum in the control sample (Table S7) (Figure 3). The relative abundance of this species in the stabilized infant gut microbial community corresponds to <0.01%. Thus, the presence of B. longum species in the control sample represents background noise that is not expected to affect downstream bioinformatic analyses.

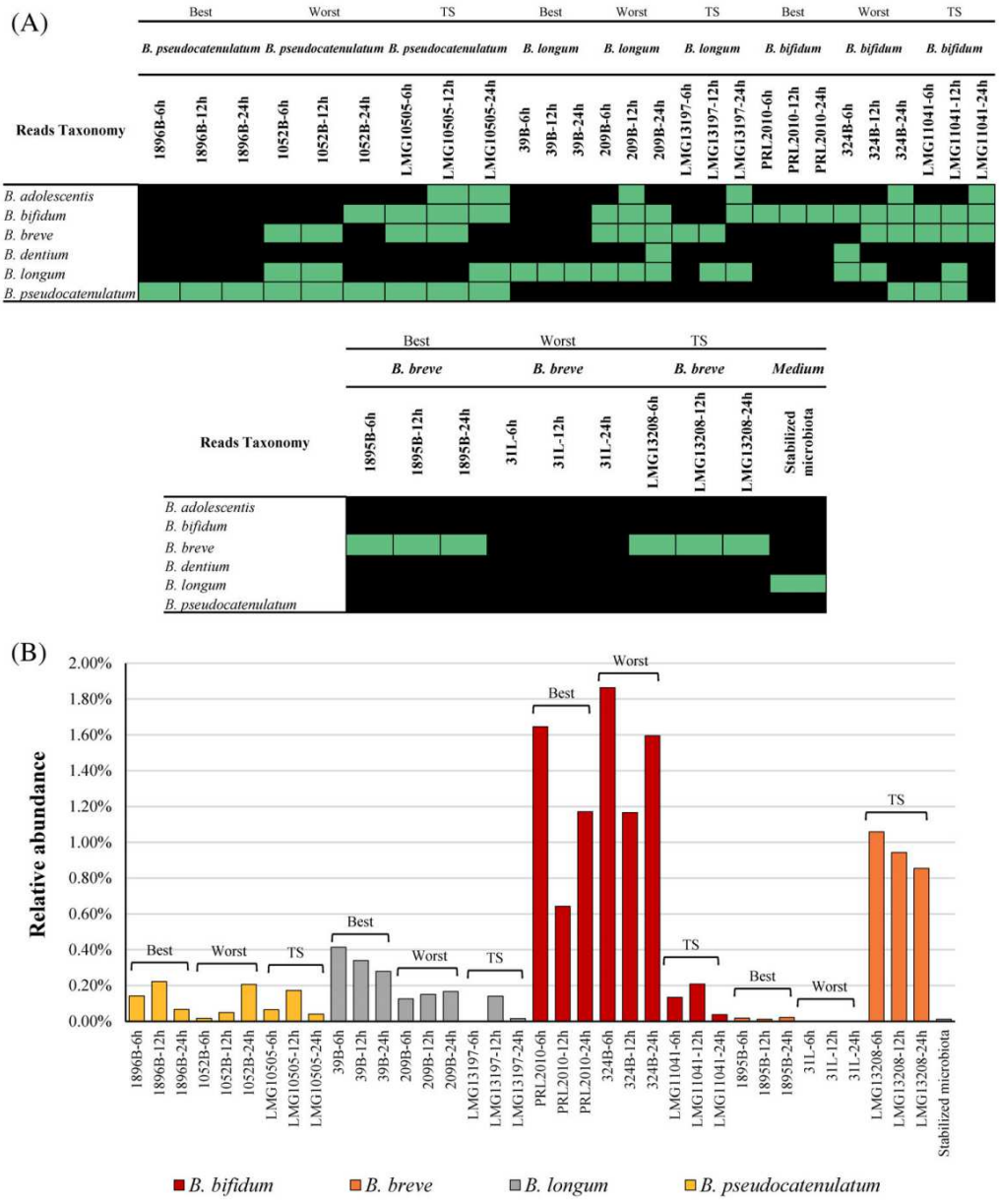


Figure 3. Detection of LocalBifRef-IV in a simulated intestinal environment. Panel (A) depicts the presence analysis of tested bifidobacterial strains with >0.01% relative abundance when grown in a simulated intestinal environment. Panel (B) reports the relative abundance reached by

the same bifidobacterial strains after 6, 12 and 24 h of fermentation. Each strain was given the adjective of worst (suboptimal candidate), best (LocalRefBif-IV) and TS (former type strain).

Remarkably, taxonomic profiles of the obtained microbial cultures inoculated with LocalRefBif-IV strains revealed the ability of these selected bifidobacterial strains to actively grow, as evidenced by a relative average abundance of 0.14% for *B. pseudocatenulatum*, 0.34% for *B. longum* and 1.15% for *B. bifidum* between the three tested time points, with the exception of *B. breve* 1895B with only a 0.02% of average relative abundance, values that correspond to the average relative abundance of the *Bifidobacterium* genus in the human gut microbiota, typically <2% (at genus level) (Do Nam et al., 2011; Odamaki et al., 2016) (Table S7) (Figure 3). In contrast, the former bifidobacterial type strains of these species included in the analysis showed lower colonization performances, as evidenced by a relative average abundance of 0.093% for *B. pseudocatenulatum*, 0.052% for *B. longum* and 0.13% for *B. bifidum* between the three tested time points (Tables S6 and S7) (Figure 3). An exception to this pattern was observed for the *B. breve* LMG13208 type strain, which showed an average relative abundance of 0.952% (Tables S6 and S7) (Figure 3). In addition, the sub-optimal model bifidobacterial strains that we used in our trials showed an average relative abundance that was somewhere between the former type strains and the optimal model strains as suggested by LocalRefBif-IV, except for sub-optimal *B. breve* 31L, which showed no growth both at 6, 12 and 24 h (Tables S6 and S7) (Figure 3).

Notably, these results confirm that the ecologic and genomic-based approach proposed for the choice of optimal microbial models allows the identification of

strains able to colonize and persist in complex microbial communities, allowing the study of the microbe–microbe interplay.

LocalRefBif-*IV* as suitable model strains for *in vitro* bifidobacteria-host interaction experiments.

The ability to interact with the host is another important feature required for an optimal reference model strain representing a commensal bacterium residing in the human gut. Thus, in order to validate the latter characteristic for the bifidobacterial strains suggested by LocalRefBif-IV, we performed in vitro trials involving human cell lines where *B. bifidum* PRL2010, *B. longum* subsp. *longum* 39B, *B. breve* 1895B and *B. pseudocatenulatum* 1896B strains were placed in contact. In addition, for comparison purposes, we included the four strains identified in our local bifidobacterial biobank as the most dissimilar to the RefBif-IS of *B. bifidum* and *B. longum* subsp. *longum*, *B. breve* and *B. pseudocatenulatum* along with the respective deposited type strains (Table S8). The cross-talk features of these strains with the human host were then inspected by transcriptomics experiments (Table S8). In this context, each of the 12 strains was incubated on co-cultured Caco2/HT29-MTX cell monolayers. Transcriptomic profiles, acquired by RNAseq experiments, were compared with those obtained from the same strains cultivated in batch on DMEM liquid media in the same incubation conditions without any contact with human cell monolayers, representing the reference conditions (Table 3) (Figure 4).

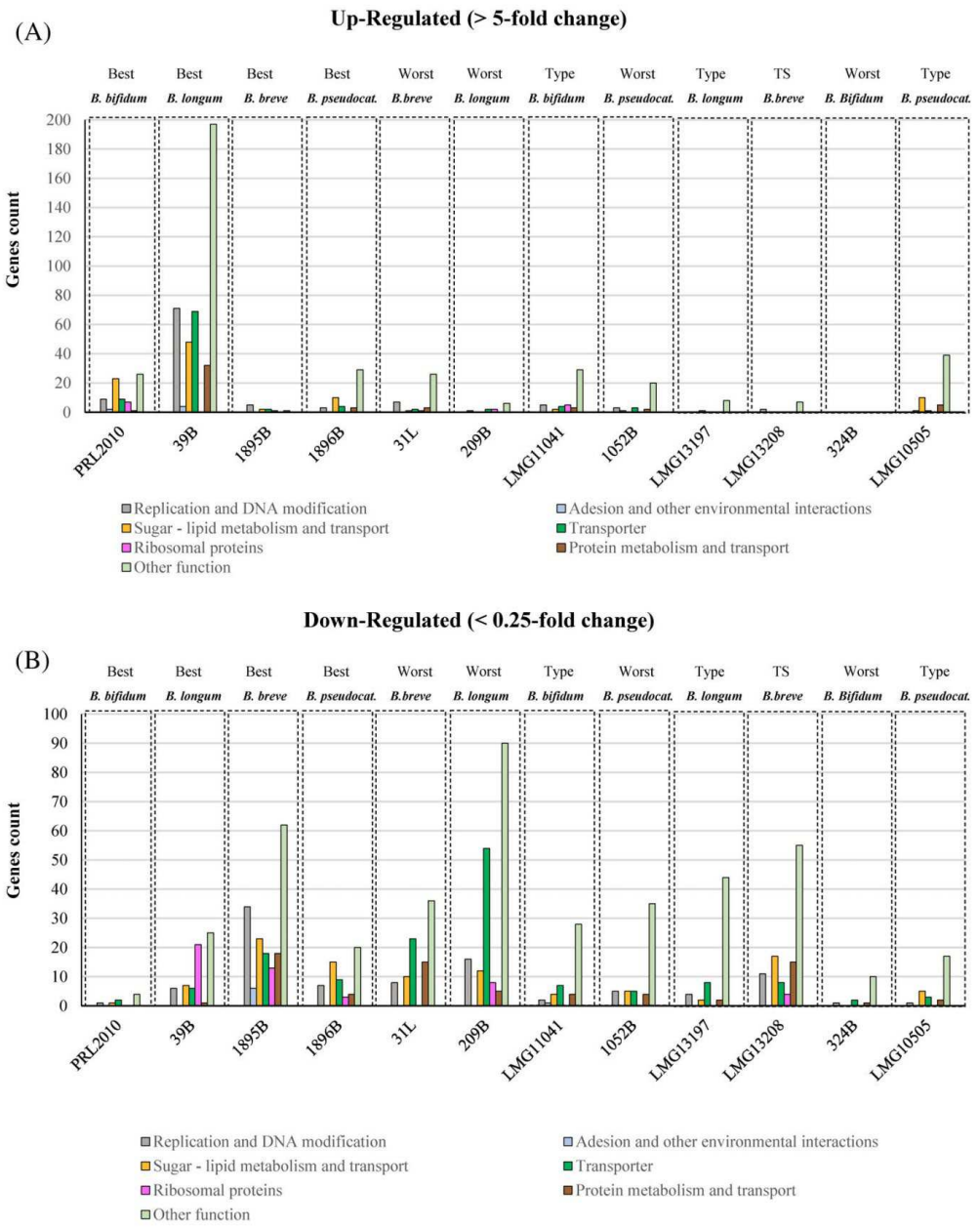


Figure 4. Transcriptomics of LocalRefBif-IV strains incubated on Caco2/HT29-MTX cell monolayers. Panel (A) reports a selection of upregulated (>5-fold change) genes, subdivided in family function, of the four selected LocalRefBif-IV strain placed in contact with Caco2/

HT29-MTX cell monolayers. Panel (B) reports a selection of down-regulated (<0.25-fold change) genes, subdivided in family function, of the four selected LocalRefBif-IV strains incubated on Caco2/HT29-MTX cell monolayers.

TABLE 3

RNA expression profiling summary

Strains growth on human cell monolayer versus control						
	Fold change					
	>3	>5	<0.25	<=0.5		
Tested strain	Gene count				Up (>5)/down (<0.25)	Type
PRL2010	207	77	8	79	9.6250	Best
39B	732	421	66	183	6.3788	Best
1895B	26	11	174	420	0.0632	Best
1896B	91	49	58	303	0.8448	Best
31 L	179	40	92	265	0.4348	Worst
209B	52	11	185	591	0.0595	Worst
LMG11041	159	48	46	178	1.0435	TS
1052B	80	29	54	398	0.5370	Worst
LMG13197	61	9	60	178	0.1500	TS
LMG13208	46	9	110	438	0.0818	TS
324B	79	0	14	223	0.0000	Worst
LMG10505	187	56	28	146	2.0000	TS

Note: Best: optimal local reference strain; TS: type strain of relative species; Worst: sub-optimal local strain.

B. bifidum PRL2010 showed 77 overexpressed (fold change > 5) and only eight downregulated (fold change < 0.25) genes, resulting in extensive activation of genes involved in the metabolism of carbon sources (24) and trans-membrane transport (11), as well as replication and signalling (10) along with two genes putatively involved in extracellular adhesion (Table S8) (Table 3) (Figure 4). Furthermore, *B. longum* subsp. *longum* 39B data showed 421 up-regulated and 66 down-regulated genes, thus being extensively affected by the interaction with

human cells. In detail, *B. longum* subsp. *longum* 39B showed activation of genes encoding the biosynthetic machinery for Tad pili, 55 genes involved in the metabolism of carbohydrates and lipids along with a notable number, that is, 76, of transporter-encoding genes and 66 genes with a putative role in transcriptional regulation (Table S8). In contrast, *B. breve* 1895B and *B. pseudocatenulatum* 1896B showed a strong transcriptional downregulation of genes, with a ratio between up-regulated and down-regulated (in the test situation and compared with control) equal to 0.0632 and 0.8448, respectively (Table 3).

These results were compared with those obtained from the corresponding type strains and for the four strains locally available and identified as the most genetically diverging from RefBif-IS (Table S6). Intriguingly, the sub-optimal strain *B. bifidum* 324B showed no up-regulated genes and only 14 down-regulated genes, highlighting the limited response of this strain to contact with human cells, while *B. bifidum* type strain LMG11041 showed a total of 94 genes with a ratio of 1.01 between up-regulated and down-regulated. In contrast, LocalRefBif-IV *B. bifidum* PRL2010, with a total of 85 genes and a ratio of 9.63 between up-regulated and down-regulated, resulted to be extensively stimulated by the contact with the human cell monolayers (Table 3) (Figure 4). Similar results can be observed analysing LocalRefBif-IV *B. longum* subsp. *longum* 39B that, with a total of 489 genes and a ratio of 6.38 between up-regulated and down-regulated genes, showed a strong transcriptional activation, which is higher than the original type strain *B. longum* subsp. *longum* LMG13197 and the sub-optimal strain *B. longum* subsp. *longum* 209B (Table 3) (Figure 4). Instead, all analysed strains of *B. breve* showed a ratio between up/down regulated genes ranging between 0.06 and 0.4, highlighting a strong down-regulation of genes in *B. breve*

species when in contact with human cell monolayers (Table S8) (Table 3) (Figure 4). Altogether, these data confirm that *B. bifidum* PRL2010, *B. longum* subsp. *longum* 39B, *B. breve* 1895B and *B. pseudocatenulatum* 1896B represent suitable model strains, as expected, due to their genetic affinity with the RefBif-IS.

Remarkably, co-cultivation with human cell lines confirmed that the identification of model microbes based on ecological and genomic data allows the definition of optimal reference model strains for the investigation of the intricate network of interactions existing between microbes and their host.

Conclusions

In this study, bifidobacterial species representing the 'core' infant gut microbiota were exploited as model taxa for the development of a novel approach to determine optimal reference strains based on strain tracking and comparative genomics investigations. Based on the integration of ecological, genomic and functional data, this approach was successfully employed to identify reference strains as research models that are representative of their corresponding species. In this regard, we employed a multi-omics approach that demonstrated how the proposed RefBif-IS reference strains can be used to carry out various in vitro experiments, that are not limited to the taxonomical classification alone. Moreover, to facilitate the use of these enhanced model reference strains, a user-friendly tool named RefBifSelector was developed to enable the screening of local strain biobanks to identify the strains phylogenetically closer to the RefBif-IS references, referred to as LocalRefBif-IV.

In this context, a model microorganism should be represented by an optimal reference strain for each bacterial species in terms of containing the genetic

capabilities of the members of that bacterial species to establish within its natural ecological niche. Therefore, when we applied this genomic-approach to identify reference strains for certain key bifidobacterial species that are naturally residing in the human gut, we decided to validate the novel identified bifidobacterial type strains identified through LocalRefBif-IV strains by performing in vitro analyses of their genetic capabilities to interact with other members of the gut microbiota as well as with the human host. Moreover, comparisons of the functional results with those retrieved for the actual recognized type strains for the bifidobacterial species here assayed and additional sub-optimal candidates highlighted that the LocalRefBif-IV possess all the characteristics to be considered excellent reference strains candidates for in vivo experiments, as well as being the most genetically similar to RefBif-IS, as expected from the collected ecological and functional data.

Data availability

Raw sequences of shallow shotgun sequencing coupled with RNA sequencing data are accessible through the Bioproject PRJNA844015 on NCBI. RefBifSelector software is downloadable from the https://probiogenomics.unipr.it/cmu/web page.

Acknowledgments

The authors thank GenProbio Srl for the financial support of the Laboratory of Probiogenomics. Part of this research benefited from the High-Performance Computing (HPC) facility of the University of Parma, Italy. Douwe van Sinderen is member of The APC Microbiome Institute funded by Science Foundation

Ireland (SFI), through the Irish Government's National Development Plan (grant numbers SFI/12/RC/2273-P1 and SFI/12/RC/2273-P2). This work was financially supported by a postDoc fellowship (Bando Ricerca Finalizzata) to Giulia Alessandri. Francesca Turroni is funded by the Italian Ministry of Health through the Bando Ricerca Finalizzata (grant number GR-2018-12365988). This research has financially been supported by the Programme "FIL-Quota Incentivante" of University of Parma and co-sponsored by Fondazione Cariparma. Finally, this study was supported by Fondazione Cariparma in the frame of the project 'Parma Microbiota' with the affiliation to the MRH of Giuseppe Taurino. Open Access Funding provided by Universita degli Studi di Parma within the CRUI-CARE Agreement.

References

Agarwala, R., Barrett, T., Beck, J., Benson, D.A., Bollin, C., Bolton, E., et al. (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **46**: D8–D13.

Alessandri, G., Ossiprandi, M.C., MacSharry, J., van Sinderen, D., and Ventura, M. (2019) Bifidobacterial Dialogue With Its Human Host and Consequent Modulation of the Immune System. *Front Immunol* **10**:.

Alessandri, G., van Sinderen, D., and Ventura, M. (2021) The genus bifidobacterium: From genomics to functionality of an important component of the mammalian gut microbiota running title: Bifidobacterial adaptation to and interaction with the host. *Comput Struct Biotechnol J* **19**: 1472–1487.

Arboleya, S., Watkins, C., Stanton, C., and Ross, R.P. (2016) Gut bifidobacteria populations in human health and aging. *Front Microbiol* **7**:.

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

Bernard, G., Greenfield, P., Ragan, M.A., and Chan, C.X. (2018) k -mer Similarity, Networks of Microbial Genomes, and Taxonomic Rank . *mSystems* **3**:.

Le Blay, G., Chassard, C., Baltzer, S., and Lacroix, C. (2010) Set up of a new in vitro model to study dietary fructans fermentation in formula-fed babies. *Br J Nutr* **103**: 403–411.

Bottacini, F., Morrissey, R., Esteban-Torres, M., James, K., Van Breen, J., Dikareva, E., et al. (2018) Comparative genomics and genotype-phenotype associations in Bifidobacterium breve. *Sci Rep* **8**:.

Bottacini, F., Van Sinderen, D., and Ventura, M. (2017) Omics of bifidobacteria: Research and insights into their health-promoting activities. *Biochem J* **474**: 4137–4152.

Cinquin, C., Le Blay, G., Fliss, I., and Lacroix, C. (2004) Immobilization of infant fecal microbiota and utilization in an in vitro colonic fermentation model. *Microb Ecol* **48**: 128–138.

van Dijk, L.R., Walker, B.J., Straub, T.J., Worby, C.J., Grote, A., Schreiber IV, H.L., et al. (2021) StrainGE: A toolkit to track and characterize low-abundance strains in complex microbial communities. *bioRxiv* 2021.02.14.431013.

Doo, E.H., Chassard, C., Schwab, C., and Lacroix, C. (2017) Effect of dietary nucleosides and yeast extracts on composition and metabolic activity of infant gut microbiota in PolyFermS colonic fermentation models. *FEMS Microbiol Ecol* **93**:.

Duranti, S., Lugli, G.A., Mancabelli, L., Armanini, F., Turroni, F., James, K., et al. (2017) Maternal inheritance of bifidobacterial communities and bifidophages in infants through vertical transmission. *Microbiome* **5**:.

Hidalgo-Cantabrana, C., Delgado, S., Ruiz, L., Ruas-Madiedo, P., Sánchez, B., and Margolles, A. (2017) Bifidobacteria and Their Health-Promoting Effects. *Microbiol Spectr* **5**:.

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**:.

Holland, N.N. (2019) Preface to the 1975 Edition. *Dyn Lit Response* XIII–XVI.

Kyrpides, N.C., Hugenholtz, P., Eisen, J.A., Woyke, T., Göker, M., Parker, C.T., et al. (2014) Genomic Encyclopedia of Bacteria and Archaea: Sequencing a Myriad of Type Strains. *PLoS Biol* **12**:.

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

Laursen, M.F., Sakanaka, M., von Burg, N., Mörbe, U., Andersen, D., Moll, J.M., et al. (2021) Bifidobacterium species associated with breastfeeding produce aromatic lactic acids in the infant gut. *Nat Microbiol* **6**: 1367.

Leinonen, R., Sugawara, H., and Shumway, M. (2011) The sequence read archive. *Nucleic Acids Res* **39**: D19.

Liang, X., Sha, Q., Rho, Y., and Zhang, S. (2018) A hierarchical clustering method for dimension reduction in joint analysis of multiple phenotypes. *Genet Epidemiol* **42**: 344–353.

Lugli, G.A., Alessandri, G., Milani, C., Viappiani, A., Fontana, F., Tarracchini, C., et al. (2021) Genetic insights into the dark matter of the mammalian gut microbiota through targeted genome reconstruction. *Environ Microbiol* **23**: 3294–3305.

Lugli, G.A., Duranti, S., Albert, K., Mancabelli, L., Napoli, S., Viappiani, A., et al. (2019) Unveiling genomic diversity among members of the species Bifidobacterium pseudolongum, a widely distributed gut commensal of the animal kingdom. *Appl Environ Microbiol* **85**:.

Lugli, G.A., Duranti, S., Milani, C., Mancabelli, L., Turroni, F., van Sinderen, D., and Ventura, M. (2019) Uncovering bifidobacteria via targeted sequencing of the mammalian gut microbiota. *Microorganisms* **7**:.

Lugli, G.A., Mancino, W., Milani, C., Duranti, S., Mancabelli, L., Napoli, S., et al. (2019) Dissecting the evolutionary development of the species bifidobacterium animalis through comparative genomics analyses. *Appl Environ Microbiol* **85**:.

Lugli, G.A., Milani, C., Duranti, S., Alessandri, G., Turroni, F., Mancabelli, L., et al. (2019) Isolation of novel gut bifidobacteria using a combination of metagenomic and cultivation approaches. *Genome Biol* **20**:.

Lugli, G.A., Tarracchini, C., Alessandri, G., Milani, C., Mancabelli, L., Turroni, F., et al. (2020) Decoding the genomic variability among members of the bifidobacterium dentium species. *Microorganisms* **8**: 1–18.

Macfarlane, G.T., Macfarlane, S., and Gibson, G.R. (1998a) Validation of a three-stage

compound continuous culture system for investigating the effect of retention time on the ecology and metabolism of bacteria in the human colon. *Microb Ecol* **35**: 180–187.

Macfarlane, G.T., Macfarlane, S., and Gibson, G.R. (1998b) Validation of a three-stage compound continuous culture system for investigating the effect of retention time on the ecology and metabolism of bacteria in the human colon. *Microb Ecol* **35**: 180–187.

Milani, C., Alessandri, G., Mancabelli, L., Mangifesta, M., Lugli, G.A., Viappiani, A., et al. (2020) Multi-omics Approaches To Decipher the Impact of Diet and Host Physiology on the Mammalian Gut Microbiome. *Appl Environ Microbiol* **86**:.

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

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

Milani, C., Mancabelli, L., Lugli, G.A., Duranti, S., Turroni, F., Ferrario, C., et al. (2015) Exploring vertical transmission of bifidobacteria from mother to child. *Appl Environ Microbiol* **81**: 7078–7087.

Milani, C., Turroni, F., Duranti, S., Lugli, G.A., Mancabelli, L., Ferrario, C., et al. (2016) Genomics of the genus Bifidobacterium reveals species-specific adaptation to the glycan-rich gut environment. *Appl Environ Microbiol* **82**: 980–991.

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

Nam, Y. Do, Jung, M.J., Roh, S.W., Kim, M.S., and Bae, J.W. (2011) Comparative analysis of korean human gut microbiota by barcoded pyrosequencing. *PLoS One* **6**: 22109.

Odamaki, T., Kato, K., Sugahara, H., Hashikura, N., Takahashi, S., Xiao, J.Z., et al. (2016) Age-related changes in gut microbiota composition from newborn to centenarian: A cross-sectional study. *BMC Microbiol* **16**: 1–12.

OriginLab (2020) Origin 9.7.0.188: Scientific Data Analysis and Graphing Software. *Orig Orig Introd* .

Parker, C.T., Tindall, B.J., and Garrity, G.M. (2019) International code of nomenclature of

Prokaryotes. *Int J Syst Evol Microbiol* **69**: S1.

Pham, V.T., Chassard, C., Rifa, E., Braegger, C., Geirnaert, A., Rocha Martin, V.N., and Lacroix, C. (2019) Lactate Metabolism Is Strongly Modulated by Fecal Inoculum, pH, and Retention Time in PolyFermS Continuous Colonic Fermentation Models Mimicking Young Infant Proximal Colon. *mSystems* **4**:.

Rivière, A., Selak, M., Lantin, D., Leroy, F., and De Vuyst, L. (2016) Bifidobacteria and butyrate-producing colon bacteria: Importance and strategies for their stimulation in the human gut. *Front Microbiol* **7**:.

Saturio, S., Nogacka, A.M., Alvarado-jasso, G.M., Salazar, N., de los Reyes-Gavilán, C.G., Gueimonde, M., and Arboleya, S. (2021) Role of Bifidobacteria on Infant Health. *Microorganisms* **9**:.

Tarracchini, C., Milani, C., Lugli, G.A., Mancabelli, L., Fontana, F., Alessandri, G., et al. (2021) Phylogenomic disentangling of the bifidobacterium longum subsp. infantis taxon. *Microb Genomics* **7**:.

Tarracchini, C., Viglioli, M., Lugli, G.A., Mancabelli, L., Fontana, F., Alessandri, G., et al. (2022) The Integrated Probiotic Database: a genomic compendium of bifidobacterial health-promoting strains. *Microbiome Res Reports* **1**: 9.

Turroni, F., Milani, C., Duranti, S., Ferrario, C., Lugli, G.A., Mancabelli, L., et al. (2018) Bifidobacteria and the infant gut: an example of co-evolution and natural selection. *Cell Mol Life Sci* **75**: 103–118.

Turroni, F., Milani, C., Duranti, S., Mahony, J., van Sinderen, D., and Ventura, M. (2018) Glycan Utilization and Cross-Feeding Activities by Bifidobacteria. *Trends Microbiol* **26**: 339–350.

Turroni, F., Milani, C., Duranti, S., Mancabelli, L., Mangifesta, M., Viappiani, A., et al. (2016) Deciphering bifidobacterial-mediated metabolic interactions and their impact on gut microbiota by a multi-omics approach. *ISME J* **10**: 1656–1668.

Turroni, F., Milani, C., Ventura, M., and van Sinderen, D. (2022) The human gut microbiota during the initial stages of life: insights from bifidobacteria. *Curr Opin Biotechnol* **73**: 81–87.

Turroni, F., van Sinderen, D., and Ventura, M. (2021) Bifidobacteria: insights into the biology

of a key microbial group of early life gut microbiota. *Microbiome Res Reports* **1**: 2.

Turroni, F., Taverniti, V., Ruas-Madiedo, P., Duranti, S., Guglielmetti, S., Lugli, G.A., et al. (2014) Bifidobacterium bifidum PRL2010 Modulates the Host Innate Immune Response. *Appl Environ Microbiol* **80**: 730–740.

Zihler Berner, A., Fuentes, S., Dostal, A., Payne, A.N., Vazquez Gutierrez, P., Chassard, C., et al. (2013) Novel Polyfermentor intestinal model (PolyFermS) for controlled ecological studies: validation and effect of pH. *PLoS One* **8**:.

Chapter 8

General Conclusion

Insight into the impact of lifestyle, geographical localization, and dairy consumption on the shaping of human gut microbiota

The study of the gut microbiota is a field of research that is becoming increasingly popular mainly due to its connections with multiple aspects of human health[126]. Therefore, it is of paramount importance to explore the factors that shape the gut microbiota compositions and its ability to maintain a homeostatic balance.

Advancements in metagenomic approaches, coupled with the progress in sequencing techniques, have allowed a detailed exploration of the microbial compositions in an increasing number of subjects. The resulting expansion of databases and increased availability of DNA sequencing data allowed the association of a large pool of metadata (such as diet, health status, stress level, and lifestyle) with different microbial profiles, opening up a new avenue of research in applied sciences. Notably, the taxonomy classification of bacteria from the phylum down to the species level has undergone continuous revisions in recent years, primarily driven by extensive comprehensive comparative genomics studies based on the growing number of genomes deposited in public databases. Hence, this thesis employed the same taxonomy used in reported scientific articles to ensure uniformity when recalling bacterial taxa.

Drastic changes in the gut microbiota composition of competitive athletes can result in severe consequences on their athletic performances. Notably, athletes in optimal psychophysical condition appear to have a stable gut microbiota capable of enhancing athletic performance. So, taking advantage of public sequencing data of fecal samples from healthy athletic and non-athletic subjects, we assessed the effects of a physically active lifestyle versus a sedentary one. So, through metagenomics analyses, we delve into the effects of lifestyles associated with competitive sports on healthy adults' gut microbiota (chapter 3, chapter 4). The study, conducted on 418 metagenomic samples, which included athletes, sedentary individuals, and moderate athletes, revealed a strong link between gut microbial profiles and athletic-related lifestyle. Specifically, a consistent microbial pattern was observed, with SCFA microbial producers such as *Faecalibacterium*, *Eubacterium*, *Blautia*, and *Ruminococcus* being prominently involved (chapter 3). Furthermore, we discovered two clusters of genes encoding for enzymes correlating with different microbial compositions. In detail, one cluster was closely associated with sedentary individuals, while the other cluster was associated with athletes' subjects (chapter 4). Competitive athletes' lifestyle significantly impacts their intestinal microbiota composition, showing a well-defined microbial population with a high rate of bacterial species capable of increasing energy intake at the intestinal level, potentially influencing host performance.

As interesting as the correlation between microbial composition and athletic performance is, many other factors interest the whole human population. One among all is the geographical localization, which can be seen as a massive metadata including several variables exerting a unique pressure on the gut

microbiota stability. Indeed, the geographical-related modulatory effect differs substantially between urban and rural regions, each characterized by distinct combinations of diet, environment, hygiene, and stress levels. So, we decided to investigate how geographical origin might influence the gut microbiota composition during the early stages of life (chapter 5). The choice to analyze only the geographic impact over the gut microbiota composition of infant subjects was to remove potential confounding metadata unrelated to geographic impact, such as age. The inclusion of 1098 gut microbiota datasets of infants revealed geographically distinct microbial patterns and differences between urbanized and pre-urbanized infants. In detail, samples from pre-urbanized regions showed much lower microbial variability than those from urbanized ones. Indeed, mothers' diets in pre-urbanized areas, which are characterized by high sugar consumption and seasonal foods, can indirectly influence the infant's gut microbiota through the modulation of human milk oligosaccharides. These intriguing findings highlight the need for additional research to fully understand the complexities behind the difference in gut microbiota composition between infants from different regions.

Moreover, among the many variables that impact the worldwide gut microbiota of humans, diet is one of the most critical. In fact, long-term dietary regimens exert a significant influence on the composition of the gut microbiota in both direct and indirect ways. In detail, diet affects the food substrate that reaches the gut, allowing certain bacterial strains to thrive and shifting the overall bacterial composition. Furthermore, it has been demonstrated that the microbiota found in fermented foods, like Italian raw milk cheeses, can serve as carriers for bacteria when consumed. Consequently, understanding as much as possible regarding the

bacterial variability found in Italian raw milk cheeses may assist in predicting which bacterial species will eventually reach the human gut and influence the host's metabolism. So, we examined 128 raw milk cheeses through shotgun metagenomics from various regions of Italy to gain insight into the cheese microbiota that can be consumed through diet. We found that PDO cheeses of the same type, but produced in different dairy factories, possess distinct microbial, metabolic, and genetic characteristics. Focusing on analyzing different types of cheeses allowed the identification of five recurring microbial profiles that are common in Italian fermented products. Several bacterial taxa were found in large part of Italian PDO cheeses, including *Lactococcus lactis*, various *Lactobacillus* species, *Streptococcus thermophilus*, *Leuconostoc mesenteroides*, and *Bifidobacterium mongoliense*. These findings can be used to better understand the modulatory effect that dairy food consumption may have on the final consumer, as well as to develop future microbial starters in industrial production for producing cheeses with excellent organoleptic qualities.

Finally, in order to corroborate our *in silico* analyses about the different modulatory agents of the human gut microbiota, using *in vitro* experiments (e.g., bioreactor model simulating the human gut or human cell monolayers), we developed a novel bioinformatic pipelines allowing the identification of valuable model bacterium of the main members of the human gut microbiota. Specifically, we used the *Bifidobacterium* genus as a test case, which represents the dominant bacterial group of the infant gut microbiota [62,127,128]. Such approach led to successfully identifying model bifidobacterial strains for the *B. bifidum*, *B. adolescentis*, *B. longum* and *B. breve* taxa, whose interactomics features with the human host were assessed using an in vitro host model. The developed strategy

represents an important achievement that fulfills the current aim of the Applied Microbiology, moving from searching which microorganism is present to a more intriguing and challenging task represented by exploring what they can do. Thus, the identification of novel model bacterial strains that well represent the genetic and ecological properties of a specific microbial taxon will be of paramount importance in order to get insights into the functionality of the human gut microbiota.

References

1 Berg, G. *et al.* Correction to: Microbiome definition re-visited: old concepts and new challenges. *Microbiome* **8**, 119, doi:10.1186/s40168-020-00905-x (2020).

2 Grice, E. A. & Segre, J. A. The skin microbiome. *Nat Rev Microbiol* **9**, 244-253, doi:10.1038/nrmicro2537 (2011).

3 Kwong, W. K. & Moran, N. A. Evolution of host specialization in gut microbes: the bee gut as a model. *Gut Microbes* **6**, 214-220, doi:10.1080/19490976.2015.1047129 (2015).

4 Deutschmann, I. M. *et al.* Disentangling microbial networks across pelagic zones in the tropical and subtropical global ocean. *Nat Commun* **15**, 126, doi:10.1038/s41467-023-44550-y (2024).

5 Bahram, M. *et al.* Structure and function of the global topsoil microbiome. *Nature* **560**, 233-237, doi:10.1038/s41586-018-0386-6 (2018).

6 Fan, Y. & Pedersen, O. Gut microbiota in human metabolic health and disease. *Nat Rev Microbiol* **19**, 55-71, doi:10.1038/s41579-020-0433-9 (2021).

7 Hertli, S. & Zimmermann, P. Molecular interactions between the intestinal microbiota and the host. *Mol Microbiol* **117**, 1297-1307, doi:10.1111/mmi.14905 (2022).

8 Sommer, F., Anderson, J. M., Bharti, R., Raes, J. & Rosenstiel, P. The resilience of the intestinal microbiota influences health and disease. *Nat Rev Microbiol* **15**, 630-638, doi:10.1038/nrmicro.2017.58 (2017).

9 Talapko, J. *et al.* Homeostasis and Dysbiosis of the Intestinal Microbiota: Comparing Hallmarks of a Healthy State with Changes in Inflammatory Bowel Disease. *Microorganisms* **10**, doi:10.3390/microorganisms10122405 (2022).

10 Hou, K. *et al.* Microbiota in health and diseases. *Signal Transduct Target Ther* **7**, 135, doi:10.1038/s41392-022-00974-4 (2022).

11 Zheng, D., Liwinski, T. & Elinav, E. Interaction between microbiota and immunity in health and disease. *Cell Res* **30**, 492-506, doi:10.1038/s41422-020-0332-7 (2020).

12 Takiishi, T., Fenero, C. I. M. & Camara, N. O. S. Intestinal barrier and gut microbiota: Shaping our immune responses throughout life. *Tissue Barriers* **5**, e1373208, doi:10.1080/21688370.2017.1373208 (2017).

13 Marrs, T. *et al.* Gut microbiota development during infancy: Impact of introducing allergenic foods. *J Allergy Clin Immunol* **147**, 613-621 e619, doi:10.1016/j.jaci.2020.09.042 (2021).

14 Wei, L., Singh, R., Ro, S. & Ghoshal, U. C. Gut microbiota dysbiosis in functional gastrointestinal disorders: Underpinning the symptoms and pathophysiology. *JGH Open* **5**, 976-987, doi:10.1002/jgh3.12528 (2021).

15 Carding, S., Verbeke, K., Vipond, D. T., Corfe, B. M. & Owen, L. J. Dysbiosis of the gut microbiota in disease. *Microb Ecol Health Dis* **26**, 26191, doi:10.3402/mehd.v26.26191 (2015).

16 Hrncir, T. Gut Microbiota Dysbiosis: Triggers, Consequences, Diagnostic and Therapeutic Options. *Microorganisms* **10**, doi:10.3390/microorganisms10030578 (2022).

17 Kaur, H., Ali, S. A. & Yan, F. Interactions between the gut microbiota-derived functional factors and intestinal epithelial cells - implication in the microbiota-host mutualism. *Front Immunol* **13**, 1006081, doi:10.3389/fimmu.2022.1006081 (2022).

18 Fusco, W. *et al.* Short-Chain Fatty-Acid-Producing Bacteria: Key Components of the Human Gut Microbiota. *Nutrients* **15**, doi:10.3390/nu15092211 (2023).

19 Furusawa, Y. *et al.* Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature* **504**, 446-450, doi:10.1038/nature12721 (2013).

20 Capozzi, V., Russo, P., Duenas, M. T., Lopez, P. & Spano, G. Lactic acid bacteria producing B-group vitamins: a great potential for functional cereals products. *Appl Microbiol Biotechnol* **96**, 1383-1394, doi:10.1007/s00253-012-4440-2 (2012).

21 Zimmermann, M., Zimmermann-Kogadeeva, M., Wegmann, R. & Goodman, A. L. Mapping human microbiome drug metabolism by gut bacteria and their genes. *Nature* **570**, 462-467, doi:10.1038/s41586-019-1291-3 (2019).

22 Louis, P., Hold, G. L. & Flint, H. J. The gut microbiota, bacterial metabolites and colorectal cancer. *Nat Rev Microbiol* **12**, 661-672, doi:10.1038/nrmicro3344 (2014).

23 Chen, Y. E., Fischbach, M. A. & Belkaid, Y. Skin microbiota-host interactions. *Nature* **553**, 427-436, doi:10.1038/nature25177 (2018).

24 Canakis, A., Haroon, M. & Weber, H. C. Irritable bowel syndrome and gut microbiota. *Curr Opin Endocrinol Diabetes Obes* **27**, 28-35, doi:10.1097/MED.0000000000000523 (2020).

25 Shaikh, S. D., Sun, N., Canakis, A., Park, W. Y. & Weber, H. C. Irritable Bowel Syndrome and the Gut Microbiome: A Comprehensive Review. *J Clin Med* **12**, doi:10.3390/jcm12072558 (2023).

26 Schirmer, M., Garner, A., Vlamakis, H. & Xavier, R. J. Microbial genes and pathways in inflammatory bowel disease. *Nat Rev Microbiol* **17**, 497-511, doi:10.1038/s41579-019-0213-6 (2019).

27 Actis, G. C., Pellicano, R., Fagoonee, S. & Ribaldone, D. G. History of Inflammatory Bowel Diseases. *J Clin Med* **8**, doi:10.3390/jcm8111970 (2019).

28 Bouma, G. & Strober, W. The immunological and genetic basis of inflammatory bowel disease. *Nat Rev Immunol* **3**, 521-533, doi:10.1038/nri1132 (2003).

29 Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655-662, doi:10.1038/s41586-019-1237-9 (2019).

30 Klunemann, M. *et al.* Bioaccumulation of therapeutic drugs by human gut bacteria. *Nature* **597**, 533-538, doi:10.1038/s41586-021-03891-8 (2021).

31 Brennan, C. A. & Garrett, W. S. Fusobacterium nucleatum - symbiont, opportunist and oncobacterium. *Nat Rev Microbiol* **17**, 156-166, doi:10.1038/s41579-018-0129-6 (2019).

32 Chen, S. *et al.* Fusobacterium nucleatum reduces METTL3-mediated m(6)A modification and contributes to colorectal cancer metastasis. *Nat Commun* **13**, 1248, doi:10.1038/s41467-022-28913-5 (2022).

33 Maier, L. *et al.* Unravelling the collateral damage of antibiotics on gut bacteria. *Nature* **599**, 120-124, doi:10.1038/s41586-021-03986-2 (2021).

34 Jakobsson, H. E. *et al.* Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. *PLoS One* **5**, e9836, doi:10.1371/journal.pone.0009836 (2010).

35 MacPherson, C. W. *et al.* Gut Bacterial Microbiota and its Resistome Rapidly Recover to Basal State Levels after Short-term Amoxicillin-Clavulanic Acid Treatment in Healthy Adults. *Sci Rep* **8**, 11192, doi:10.1038/s41598-018-29229-5 (2018).

36 Maier, L. *et al.* Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555**, 623-628, doi:10.1038/nature25979 (2018).

37 Zmora, N., Suez, J. & Elinav, E. You are what you eat: diet, health and the gut microbiota. *Nat Rev Gastroenterol Hepatol* **16**, 35-56, doi:10.1038/s41575-018-0061-2 (2019).

38 Kolodziejczyk, A. A., Zheng, D. & Elinav, E. Diet-microbiota interactions and personalized nutrition. *Nat Rev Microbiol* **17**, 742-753, doi:10.1038/s41579-019-0256-8 (2019).

39 Makki, K., Deehan, E. C., Walter, J. & Backhed, F. The Impact of Dietary Fiber on Gut Microbiota in Host Health and Disease. *Cell Host Microbe* **23**, 705-715, doi:10.1016/j.chom.2018.05.012 (2018).

40 Xu, A. A. *et al.* Dietary Fatty Acid Intake and the Colonic Gut Microbiota in Humans. *Nutrients* **14**, doi:10.3390/nu14132722 (2022).

41 Wu, S. *et al.* Effect of Dietary Protein and Processing on Gut Microbiota-A Systematic Review. *Nutrients* **14**, doi:10.3390/nu14030453 (2022).

42 Flynn, J. M., Niccum, D., Dunitz, J. M. & Hunter, R. C. Evidence and Role for Bacterial Mucin Degradation in Cystic Fibrosis Airway Disease. *PLoS Pathog* **12**, e1005846, doi:10.1371/journal.ppat.1005846 (2016).

43 Al Nabhani, Z. *et al.* A Weaning Reaction to Microbiota Is Required for Resistance to Immunopathologies in the Adult. *Immunity* **50**, 1276-1288 e1275, doi:10.1016/j.immuni.2019.02.014 (2019).

44 de Muinck, E. J. & Trosvik, P. Individuality and convergence of the infant gut microbiota during the first year of life. *Nat Commun* **9**, 2233, doi:10.1038/s41467-018-04641-7 (2018).

45 Hugenholtz, F. *et al.* Feasibility of Metatranscriptome Analysis from Infant Gut Microbiota: Adaptation to Solid Foods Results in Increased Activity of Firmicutes at Six Months. *Int J Microbiol* **2017**, 9547063, doi:10.1155/2017/9547063 (2017).

46 Vatanen, T. *et al.* A distinct clade of Bifidobacterium longum in the gut of Bangladeshi children thrives during weaning. *Cell* **185**, 4280-4297 e4212, doi:10.1016/j.cell.2022.10.011 (2022).

47 Cuamatzin-Garcia, L. *et al.* Traditional Fermented Foods and Beverages from around the World and Their Health Benefits. *Microorganisms* **10**, doi:10.3390/microorganisms10061151 (2022).

48 Kok, C. R. & Hutkins, R. Yogurt and other fermented foods as sources of health-promoting bacteria. *Nutr Rev* **76**, 4-15, doi:10.1093/nutrit/nuy056 (2018).

49 Sharma, H., Ozogul, F., Bartkiene, E. & Rocha, J. M. Impact of lactic acid bacteria and their metabolites on the techno-functional properties and health benefits of fermented dairy products. *Crit Rev Food Sci Nutr* **63**, 4819-4841, doi:10.1080/10408398.2021.2007844 (2023).

50 Mannaa, M., Han, G., Seo, Y. S. & Park, I. Evolution of Food Fermentation Processes and the Use of Multi-Omics in Deciphering the Roles of the Microbiota. *Foods* **10**, doi:10.3390/foods10112861 (2021).

51 Giraffa, G. Studying the dynamics of microbial populations during food fermentation. *FEMS Microbiol Rev* **28**, 251-260, doi:10.1016/j.femsre.2003.10.005 (2004).

52 Abedi, E. & Hashemi, S. M. B. Lactic acid production - producing microorganisms and substrates sources-state of art. *Heliyon* **6**, e04974, doi:10.1016/j.heliyon.2020.e04974 (2020).

53 Mani-Lopez, E., Arrioja-Breton, D. & Lopez-Malo, A. The impacts of antimicrobial and antifungal activity of cell-free supernatants from lactic acid bacteria in vitro and foods. *Compr Rev Food Sci Food Saf* **21**, 604-641, doi:10.1111/1541-4337.12872 (2022).

54 Fugaban, J. I. I., Jung, E. S., Todorov, S. D. & Holzapfel, W. H. Evaluation of Antifungal Metabolites Produced by Lactic Acid Bacteria. *Probiotics Antimicrob Proteins* **15**, 1447-1463, doi:10.1007/s12602-022-09995-5 (2023).

55 Daniel, N. *et al.* Gut microbiota and fermentation-derived branched chain hydroxy acids mediate health benefits of yogurt consumption in obese mice. *Nat Commun* **13**, 1343, doi:10.1038/s41467-022-29005-0 (2022).

56 Shiby, V. K. & Mishra, H. N. Fermented milks and milk products as functional foods--a review. *Crit Rev Food Sci Nutr* **53**, 482-496, doi:10.1080/10408398.2010.547398 (2013).

57 Garbacz, K. Anticancer activity of lactic acid bacteria. *Semin Cancer Biol* **86**, 356-366, doi:10.1016/j.semcancer.2021.12.013 (2022).

58 Dominguez-Bello, M. G., Godoy-Vitorino, F., Knight, R. & Blaser, M. J. Role of the microbiome in human development. *Gut* **68**, 1108-1114, doi:10.1136/gutjnl-2018-317503 (2019).

59 Mohr, A. E. *et al.* The athletic gut microbiota. *J Int Soc Sports Nutr* **17**, 24, doi:10.1186/s12970-020-00353-w (2020).

60 Campaniello, D. *et al.* How Diet and Physical Activity Modulate Gut Microbiota: Evidence, and Perspectives. *Nutrients* **14**, doi:10.3390/nu14122456 (2022).

61 Odamaki, T. *et al.* Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. *BMC Microbiol* **16**, 90, doi:10.1186/s12866-016-0708-5 (2016).

62 Milani, C. *et al.* The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. *Microbiol Mol Biol Rev* **81**, doi:10.1128/MMBR.00036-17 (2017).

63 Saturio, S. *et al.* Early-Life Development of the Bifidobacterial Community in the Infant Gut. *Int J Mol Sci* **22**, doi:10.3390/ijms22073382 (2021).

64 Ho, N. T. *et al.* Meta-analysis of effects of exclusive breastfeeding on infant gut microbiota across populations. *Nat Commun* **9**, 4169, doi:10.1038/s41467-018-06473-x (2018).

65 Mancabelli, L. *et al.* Multi-population cohort meta-analysis of human intestinal microbiota in early life reveals the existence of infant community state types (ICSTs). *Comput Struct Biotechnol J* **18**, 2480-2493, doi:10.1016/j.csbj.2020.08.028 (2020).

66 De Filippis, F. *et al.* Specific gut microbiome signatures and the associated pro-inflamatory functions are linked to pediatric allergy and acquisition of immune tolerance. *Nat Commun* **12**, 5958, doi:10.1038/s41467-021-26266-z (2021).

67 Mohr, A. E., Ahern, M. M., Sears, D. D., Bruening, M. & Whisner, C. M. Gut microbiome diversity, variability, and latent community types compared with shifts in body weight during the freshman year of college in dormitory-housed adolescents. *Gut Microbes* **15**, 2250482, doi:10.1080/19490976.2023.2250482 (2023).

68 Faith, J. J. *et al.* The long-term stability of the human gut microbiota. *Science* **341**, 1237439, doi:10.1126/science.1237439 (2013).

69 Ragonnaud, E. & Biragyn, A. Gut microbiota as the key controllers of "healthy" aging of elderly people. *Immun Ageing* **18**, 2, doi:10.1186/s12979-020-00213-w (2021).

70 Salazar, N., Valdes-Varela, L., Gonzalez, S., Gueimonde, M. & de Los Reyes-Gavilan, C. G. Nutrition and the gut microbiome in the elderly. *Gut Microbes* **8**, 82-97, doi:10.1080/19490976.2016.1256525 (2017).

71 Scepanovic, P. *et al.* A comprehensive assessment of demographic, environmental, and host genetic associations with gut microbiome diversity in healthy individuals. *Microbiome* **7**, 130, doi:10.1186/s40168-019-0747-x (2019).

72 Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210-215, doi:10.1038/nature25973 (2018).

73 Pausan, M. R., Blohs, M., Mahnert, A. & Moissl-Eichinger, C. The sanitary indoor environment-a potential source for intact human-associated anaerobes. *NPJ Biofilms Microbiomes* **8**, 44, doi:10.1038/s41522-022-00305-z (2022).

74 Browne, H. P., Neville, B. A., Forster, S. C. & Lawley, T. D. Transmission of the gut microbiota: spreading of health. *Nat Rev Microbiol* **15**, 531-543, doi:10.1038/nrmicro.2017.50 (2017).

75 Gacesa, R. *et al.* Environmental factors shaping the gut microbiome in a Dutch population. *Nature* **604**, 732-739, doi:10.1038/s41586-022-04567-7 (2022).

76 Tasnim, N., Abulizi, N., Pither, J., Hart, M. M. & Gibson, D. L. Linking the Gut Microbial Ecosystem with the Environment: Does Gut Health Depend on Where We Live? *Front Microbiol* **8**, 1935, doi:10.3389/fmicb.2017.01935 (2017).

77 Afshari, R. *et al.* New insights into cheddar cheese microbiota-metabolome relationships revealed by integrative analysis of multi-omics data. *Sci Rep* **10**, 3164, doi:10.1038/s41598-020-59617-9 (2020).

78 Carpino, S. *et al.* Influence of PDO Ragusano cheese biofilm microbiota on flavour compounds formation. *Food Microbiol* **61**, 126-135, doi:10.1016/j.fm.2016.09.006 (2017).

79 Afshari, R., Pillidge, C. J., Dias, D. A., Osborn, A. M. & Gill, H. Cheesomics: the future pathway to understanding cheese flavour and quality. *Crit Rev Food Sci Nutr* **60**, 33-47, doi:10.1080/10408398.2018.1512471 (2020).

80 Anastasiou, R. *et al.* Omics Approaches to Assess Flavor Development in Cheese. *Foods* **11**, doi:10.3390/foods11020188 (2022).

81 Sola, L. *et al.* Insights on the bacterial composition of Parmigiano Reggiano Natural Whey Starter by a culture-dependent and 16S rRNA metabarcoding portrait. *Sci Rep* **12**, 17322, doi:10.1038/s41598-022-22207-y (2022).

82 Hu, Y., Zhang, L., Wen, R., Chen, Q. & Kong, B. Role of lactic acid bacteria in flavor development in traditional Chinese fermented foods: A review. *Crit Rev Food Sci Nutr* **62**, 2741-2755, doi:10.1080/10408398.2020.1858269 (2022).

83 Johnson, J., Curtin, C. & Waite-Cusic, J. The Cheese Production Facility Microbiome Exhibits Temporal and Spatial Variability. *Front Microbiol* **12**, 644828, doi:10.3389/fmicb.2021.644828 (2021).

84 Fontana, F. *et al.* Multifactorial Microvariability of the Italian Raw Milk Cheese Microbiota and Implication for Current Regulatory Scheme. *mSystems* **8**, e0106822, doi:10.1128/msystems.01068-22 (2023).

85 Marasco, R., Gazzillo, M., Campolattano, N., Sacco, M. & Muscariello, L. Isolation and Identification of Lactic Acid Bacteria from Natural Whey Cultures of Buffalo and Cow Milk. *Foods* **11**, doi:10.3390/foods11020233 (2022).

86 Heo, S. *et al.* Safety Assessment Systems for Microbial Starters Derived from Fermented Foods. *J Microbiol Biotechnol* **32**, 1219-1225, doi:10.4014/jmb.2207.07047 (2022).

87 Lucchini, R. *et al.* Contribution of natural milk culture to microbiota, safety and hygiene of raw milk cheese produced in alpine malga. *Ital J Food Saf* **7**, 6967, doi:10.4081/ijfs.2018.6967 (2018).

88 Lugli, G. A. *et al.* Genetic insights into the dark matter of the mammalian gut microbiota through targeted genome reconstruction. *Environ Microbiol* **23**, 3294-3305, doi:10.1111/1462-2920.15559 (2021).

89 Allali, I. *et al.* A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. *BMC Microbiol* **17**, 194, doi:10.1186/s12866-017-1101-8 (2017).

90 Buermans, H. P. & den Dunnen, J. T. Next generation sequencing technology: Advances and applications. *Biochim Biophys Acta* **1842**, 1932-1941, doi:10.1016/j.bbadis.2014.06.015 (2014).

91 Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res* **50**, D20-D26, doi:10.1093/nar/gkab1112 (2022).

92 Orsini, M., Cuccuru, G., Uva, P. & Fotia, G. Bacterial Genomic Data Analysis in the Next-Generation Sequencing Era. *Methods Mol Biol* **1415**, 407-422, doi:10.1007/978-1-4939-3572-7_21 (2016).

93 Wright, M. H., Adelskov, J. & Greene, A. C. Bacterial DNA Extraction Using Individual Enzymes and Phenol/Chloroform Separation. *J Microbiol Biol Educ* **18**, doi:10.1128/jmbe.v18i2.1348 (2017).

94 Cristina Barbosa, S. N., Mário Gadanho, Sandra Chaves. in *Molecular Microbial Diagnostic Methods* Vol. Pathways to Implementation for the Food and Water

Industry (ed Martin D'Agostino Nigel Cook, K. Clive Thompson) Ch. Chapter 7, 135-154 (Academic Press, 2016).

95 Shi, Z. *et al.* The Effects of DNA Extraction Kits and Primers on Prokaryotic and Eukaryotic Microbial Community in Freshwater Sediments. *Microorganisms* **10**, doi:10.3390/microorganisms10061213 (2022).

96 Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**, 441-448, doi:10.1016/0022-2836(75)90213-2 (1975).

97 Cheng, C., Fei, Z. & Xiao, P. Methods to improve the accuracy of next-generation sequencing. *Front Bioeng Biotechnol* **11**, 982111, doi:10.3389/fbioe.2023.982111 (2023).

98 Wensel, C. R., Pluznick, J. L., Salzberg, S. L. & Sears, C. L. Next-generation sequencing: insights to advance clinical investigations of the microbiome. *J Clin Invest* **132**, doi:10.1172/JCI154944 (2022).

99 Behjati, S. & Tarpey, P. S. What is next generation sequencing? *Arch Dis Child Educ Pract Ed* **98**, 236-238, doi:10.1136/archdischild-2013-304340 (2013).

100 Churko, J. M., Mantalas, G. L., Snyder, M. P. & Wu, J. C. Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. *Circ Res* **112**, 1613-1623, doi:10.1161/CIRCRESAHA.113.300939 (2013).

101 Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* **39**, 1348-1365, doi:10.1038/s41587-021-01108-x (2021).

102 Xiao, T. & Zhou, W. The third generation sequencing: the advanced approach to genetic diseases. *Transl Pediatr* **9**, 163-173, doi:10.21037/tp.2020.03.06 (2020).

103 Pfeiffer, F. *et al.* Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci Rep* **8**, 10950, doi:10.1038/s41598-018-29325-6 (2018).

104 Brown, C. L. *et al.* Critical evaluation of short, long, and hybrid assembly for contextual analysis of antibiotic resistance genes in complex environmental metagenomes. *Sci Rep* **11**, 3753, doi:10.1038/s41598-021-83081-8 (2021).

105 Arikawa, K. & Hosokawa, M. Uncultured prokaryotic genomes in the spotlight: An examination of publicly available data from metagenomics and single-cell genomics. *Comput Struct Biotechnol J* **21**, 4508-4518, doi:10.1016/j.csbj.2023.09.010 (2023).

106 O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745, doi:10.1093/nar/gkv1189 (2016).

107 Milani, C. *et al.* METAnnotatorX2: a Comprehensive Tool for Deep and Shallow Metagenomic Data Set Analyses. *mSystems* **6**, e0058321, doi:10.1128/mSystems.00583-21 (2021).

108 Vincent, A. T., Derome, N., Boyle, B., Culley, A. I. & Charette, S. J. Next-generation sequencing (NGS) in the microbiological world: How to make the most of your money. *J Microbiol Methods* **138**, 60-71, doi:10.1016/j.mimet.2016.02.016 (2017).

109 Tan, S. *et al.* Uncovering the performance bottleneck of modern HPC processor with static code analyzer: a case study on Kunpeng 920. *CCF Transactions on High Performance Computing*, doi:10.1007/s42514-023-00160-0 (2023).

110 Xu, R., Rajeev, S. & Salvador, L. C. M. The selection of software and database for metagenomics sequence analysis impacts the outcome of microbial

profiling and pathogen detection. *PLoS One* **18**, e0284031, doi:10.1371/journal.pone.0284031 (2023).

111 Santamaria, M. *et al.* Reference databases for taxonomic assignment in metagenomics. *Brief Bioinform* **13**, 682-695, doi:10.1093/bib/bbs036 (2012).

112 Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590-596, doi:10.1093/nar/gks1219 (2013).

113 Clarridge, J. E., 3rd. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev* **17**, 840-862, table of contents, doi:10.1128/CMR.17.4.840-862.2004 (2004).

114 Aggarwal, D. *et al.* Optimization of high-throughput 16S rRNA gene amplicon sequencing: an assessment of PCR pooling, mastermix use and contamination. *Microb Genom* **9**, doi:10.1099/mgen.0.001115 (2023).

115 Johnson, J. S. *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun* **10**, 5029, doi:10.1038/s41467-019-13036-1 (2019).

116 Man, S. M., Kaakoush, N. O., Octavia, S. & Mitchell, H. The internal transcribed spacer region, a new tool for use in species differentiation and delineation of systematic relationships within the Campylobacter genus. *Appl Environ Microbiol* **76**, 3071-3081, doi:10.1128/AEM.02551-09 (2010).

117 Peker, N. *et al.* A Comparison of Three Different Bioinformatics Analyses of the 16S-23S rRNA Encoding Region for Bacterial Identification. *Front Microbiol* **10**, 620, doi:10.3389/fmicb.2019.00620 (2019).

118 Ghyselinck, J., Pfeiffer, S., Heylen, K., Sessitsch, A. & De Vos, P. The effect of primer choice and short read sequences on the outcome of 16S rRNA gene

based diversity studies. *PLoS One* **8**, e71360, doi:10.1371/journal.pone.0071360 (2013).

119 Bukin, Y. S. *et al.* The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Sci Data* **6**, 190007, doi:10.1038/sdata.2019.7 (2019).

120 Navas-Molina, J. A., Hyde, E. R., Sanders, J. & Knight, R. The Microbiome and Big Data. *Curr Opin Syst Biol* **4**, 92-96, doi:10.1016/j.coisb.2017.07.003 (2017).

121 Ranjan, R., Rani, A., Metwally, A., McGee, H. S. & Perkins, D. L. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem Biophys Res Commun* **469**, 967-977, doi:10.1016/j.bbrc.2015.12.083 (2016).

122 Usyk, M. *et al.* Comprehensive evaluation of shotgun metagenomics, amplicon sequencing, and harmonization of these platforms for epidemiological studies. *Cell Rep Methods* **3**, 100391, doi:10.1016/j.crmeth.2022.100391 (2023).

123 Zhou, Y., Liu, M. & Yang, J. Recovering metagenome-assembled genomes from shotgun metagenomic sequencing data: Methods, applications, challenges, and opportunities. *Microbiol Res* **260**, 127023, doi:10.1016/j.micres.2022.127023 (2022).

124 Jin, H. *et al.* Hybrid, ultra-deep metagenomic sequencing enables genomic and functional characterization of low-abundance species in the human gut microbiome. *Gut Microbes* **14**, 2021790, doi:10.1080/19490976.2021.2021790 (2022).

125 Hillmann, B. *et al.* Evaluating the Information Content of Shallow Shotgun Metagenomics. *mSystems* **3**, doi:10.1128/mSystems.00069-18 (2018).

126 Lynch, S. V. & Pedersen, O. The Human Intestinal Microbiome in Health and Disease. *N Engl J Med* **375**, 2369-2379, doi:10.1056/NEJMra1600266 (2016).

127 Turroni, F. *et al.* Diversity of bifidobacteria within the infant gut microbiota. *PLoS One* **7**, e36957, doi:10.1371/journal.pone.0036957 (2012).

128 Alessandri, G. *et al.* Exploring species-level infant gut bacterial biodiversity by meta-analysis and formulation of an optimized cultivation medium. *NPJ Biofilms Microbiomes* **8**, 88, doi:10.1038/s41522-022-00349-1 (2022).

Publications in peer-reviewed journals achieved in the course of the Ph.D.

1) Alessandri, G., Sangalli, E., Facchi, M., **Fontana, F**., Mancabelli, L., Donofrio, G., & Ventura, M. "**Metataxonomic analysis of milk microbiota in the bovine subclinical mastitis**".
(2023) FEMS microbiology ecology, 99(12). doi.org/10.1093/femsec/fiad136

2) Tarracchini, C., Alessandri, G., **Fontana, F**., Rizzo, S. M., Lugli, G. A., Bianchi, M. G., Mancabelli, L., Longhi, G., Argentini, C., Vergna, L. M., Anzalone, R., Viappiani, A., Turroni, F., Taurino, G., Chiu, M., Arboleya, S., Gueimonde, M., Bussolati, O., van Sinderen, D., Milani, C., … Ventura, M. "**Genetic strategies for sex-biased persistence of gut microbes across human life**".
(2023) Nature communications, 14(1), 4220. doi.org/10.1038/s41467-023-39931-2

3) Lugli, G. A., Mancabelli, L., Milani, C., **Fontana, F**., Tarracchini, C., Alessandri, G., van Sinderen, D., Turroni, F., & Ventura, M. "**Comprehensive**

insights from composition to functional microbe-based biodiversity of the infant human gut microbiota".
(2023) NPJ biofilms and microbiomes, 9(1), 25, doi.org/10.1038/s41522-023-00392-6.

4) **Fontana, F**., Longhi, G., Tarracchini, C., Mancabelli, L., Lugli, G. A., Alessandri, G., Turroni, F., Milani, C., & Ventura, M. **"The human gut microbiome of athletes: metagenomic and metabolic insights".**
(2023) Microbiome, doi:10.1186/s40168-023-01470-9.

5) Mancabelli, L., Taurino, G., Ticinesi, A., Ciociola, T., Vacondio, F., Milani, C., **Fontana, F**., Lugli, G. A., Tarracchini, C., Alessandri, G., Viappiani, A., Bianchi, M., Nouvenne, A., Chetta, A. A., Turroni, F., Meschi, T., Mor, M., Bussolati, O., & Ventura, M. "**Disentangling the interactions between nasopharyngeal and gut microbiome and their involvement in the modulation of COVID-19 infection**".
(2023) Microbiology spectrum, 11(5), e0219423. Advance online publication. doi.org/10.1128/spectrum.02194-23

6) Tarracchini, C., Argentini, C., Alessandri, G., Lugli, G. A., Mancabelli, L., **Fontana, F**., Anzalone, R., Viappiani, A., Turroni, F., Ventura, M., & Milani, C. "**The core genome evolution of Lactobacillus crispatus as a driving force for niche competition in the human vaginal tract**".
(2023) Microbial biotechnology, 16(9), 1774–1789. doi.org/10.1111/1751-7915.14305

7) Pivetta, G., Dottori, L., **Fontana, F**., Cingolani, S., Ligato, I., Dilaghi, E., Milani, C., Ventura, M., Borro, M., Esposito, G., Annibale, B., & Lahner, E. "**Gastric**

Microbiota Gender Differences in Subjects with Healthy Stomachs and Autoimmune Atrophic Gastritis".
(2023) Microorganisms, 11(8), 1938. doi.org/10.3390/microorganisms11081938

8) Rizzo, S. M., Alessandri, G., Lugli, G. A., **Fontana, F**., Tarracchini, C., Mancabelli, L., Viappiani, A., Bianchi, M. G., Bussolati, O., van Sinderen, D., Ventura, M., & Turroni, F. "**Exploring Molecular Interactions between Human Milk Hormone Insulin and Bifidobacteria**".
(2023) Microbiology spectrum, 11(3), e0066523. https://doi.org/10.1128/spectrum.00665-23

9) Alessandri, G., **Fontana, F**., Tarracchini, C., Rizzo, S. M., Bianchi, M. G., Taurino, G., Chiu, M., Lugli, G. A., Mancabelli, L., Argentini, C., Longhi, G., Anzalone, R., Viappiani, A., Milani, C., Turroni, F., Bussolati, O., van Sinderen, D., & Ventura, M. "**Identification of a prototype human gut Bifidobacterium longum subsp. longum strain based on comparative and functional genomic approaches"**.
(2023) Frontiers in microbiology, 14, 1130592. https://doi.org/10.3389/fmicb.2023.1130592

10) ***Fontana F.**, *Longhi G., Alessandri G, Lugli GA., 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 Current Regulatory Scheme".**
(2022-2023), mSystems, doi: 10.1128/msystems.01068-22.
***, ° Equal contribution.**

11) Alessandri, G., **Fontana, F**., Mancabelli, L., Lugli, G.A., Tarracchini, C., Argentini, C., Longhi, G., Viappiani, A., Milani, C., Turroni, F., van Sinderen, D.,

Ventura, M. "**Exploring species-level infant gut bacterial biodiversity by meta-analysis and formulation of an optimized cultivation medium**".
(2022) npj Biofilms and Microbiomes, doi: 10.1038/s41522-022-00349-1.

12) **Fontana 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**".
(2022) Environmental Microbiology, doi: 10.1111/1462-2920.16205.

13) Lugli, G. A., **Fontana, F**., Tarracchini, C., Mancabelli, L., Milani, C., Turroni, F., & Ventura, M. "**Exploring the biodiversity of Bifidobacterium asteroides among honey bee microbiomes**".
(2022) Environmental microbiology, 24(12), 5666–5679. https://doi.org/10.1111/1462-2920.16223

14) Mancabelli, L., Milani, C., **Fontana, F**., Lugli, G. A., Tarracchini, C., Viappiani, A., Ciociola, T., Ticinesi, A., Nouvenne, A., Meschi, T., Turroni, F., & Ventura, M. "**Untangling the link between the human gut microbiota composition and the severity of the symptoms of the COVID-19 infection**".
(2022) Environmental microbiology, 24(12), 6453–6462. https://doi.org/10.1111/1462-2920.16201

15) Tarracchini, C., **Fontana, F**., Mancabelli, L., Lugli, G.A., Alessandri, G., Turroni, F., Ventura, M., Milani, C. "**Gut microbe metabolism of small molecules supports human development across the early stages of life**".
(2022) Frontiers in Microbiology, doi:10.3389/fmicb.2022.1006721.

16) Leonardo Mancabelli, Tecla Ciociola, Gabriele Andrea Lugli, Chiara Tarracchini, **Federico Fontana**, Alice Viappiani, Francesca Turroni, Andrea Ticinesi, Tiziana Meschi, Stefania Conti, Marco Ventura & Christian Milani. "**Guideline for the analysis of the microbial communities of the human upper airways**" (2022) Journal of Oral Microbiology, 14:1, doi:10.1080/20002297.2022.2103282.

17) Lugli, G. A., Longhi, G., 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**". (2022) Environmental microbiology, 24(9), 3912–3923. doi.org/10.1111/1462-2920.15988

18) *Tarracchini, C., ***Fontana, F.,** Lugli, G.A., Mancabelli, L., Alessandri, G., Turroni, F., Ventura, M., Milani, C. "**Investigation of the Ecological Link between Recurrent Microbial Human Gut Communities and Physical Activity**". (2022) Microbiology Spectrum, doi:10.1128/spectrum.00420-22.
*** Equal contribution.**

19) *Argentini, C., ***Fontana, F.,** Alessandri, G., Lugli, G.A., Mancabelli, L., Ossiprandi, M.C., van Sinderen, D., Ventura, M., Milani, C., Turroni, F. "**Evaluation of Modulatory Activities of Lactobacillus crispatus Strains in the Context of the Vaginal Microbiota**". (2022) Microbiology Spectrum, doi:10.1128/spectrum.02733-21.
*** Equal contribution.**

20) Alessandri, G., Lugli, G.A., Tarracchini, C., Rizzo, S.M., Argentini, C., Viappiani, A., Mancabelli, L., **Fontana, F**., Milani, C., Turroni, F., van Sinderen, D., Ventura, M. "**Disclosing the Genomic Diversity among Members of the**

Bifidobacterium Genus of Canine and Feline Origin with Respect to Those from Human".

(2022) Applied and Environmental Microbiology, doi:10.1128/aem.02038-21.

21) Lugli, G.A., Longhi, G., 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**".

(2022) Frontiers in Microbiology, doi:10.3389/fmicb.2021.790881.

22) Lugli, G.A., **Fontana, F**., Tarracchini, C., Mancabelli, L., Milani, C., Turroni, F., Ventura, M. "**Exploring the biodiversity of Bifidobacterium asteroides among honey bee microbiomes**"

(2022) Environmental Microbiology, doi:10.1111/1462-2920.16223.

23) Mancabelli, L., Milani, C., **Fontana, F**., Lugli, G.A., Tarracchini, C., Viappiani, A., Ciociola, T., Ticinesi, A., Nouvenne, A., Meschi, T., Turroni, F., Ventura, M. "**Untangling the link between the human gut microbiota composition and the severity of the symptoms of the COVID-19 infection**".

(2022) Environmental Microbiology, doi:10.1111/1462-2920.16201.

24) Lugli, G.A., Longhi, G., 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**".

(2022) Environmental Microbiology, doi:10.1111/1462-2920.15988.

25) Mancabelli, L., Milani, C., **Fontana, F.,** Lugli, G.A., Tarracchini, C., Turroni, F., van Sinderen, D., Ventura, M. "**Mapping bacterial diversity and metabolic functionality of the human respiratory tract microbiome**"
(2022) Journal of Oral Microbiology, doi:10.1080/20002297.2022.2051336.

26) Mancabelli, L., Milani, C., Anzalone, R., Alessandri, G., Lugli, G.A., Tarracchini, C., **Fontana, F.,** Turroni, F., Ventura, M. Free DNA and Metagenomics Analyses: Evaluation of Free DNA "**Inactivation Protocols for Shotgun Metagenomics Analysis of Human Biological Matrices**"
(2021) Frontiers in Microbiology, doi:10.3389/fmicb.2021.749373.

27) Tarracchini, C., Milani, C., Longhi, G., **Fontana, F.,** Mancabelli, L., Pintus, R., Andrea Lugli, G., Alessandri, G., Anzalone, R., Viappiani, A., Turroni, F., Mussap, M., Dessì, A., Marincola, F.C., Noto, A., De Magistris, A., Vincent, M., Bernasconi, S., Picaud, J.-C., Fanos, V., Ventura, M. "**Unraveling the microbiome of necrotizing enterocolitis: Insights in novel microbial and metabolomic biomarkers**"
(2021) Microbiology Spectrum, doi:10.1128/Spectrum.01176-21.

28) **Fontana, F.,** Mancabelli, L., Lugli, G.A., Taracchini, C., Alessandri, G., Longhi, G., Anzalone, R., Viappiani, A., Famo, R., Brognan, M., Micondo, K.H., 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**"
(2021) Environmental Microbiology Reports, doi:10.1111/1758-2229.12960.

29) Milani, C., Lugli, G.A., **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"
(2021) mSystems, doi:10.1128/mSystems.00583-21.

30) Lugli, G.A., Alessandri, G., Milani, C., Viappiani, A., **Fontana, F.**, Tarracchini, C., Mancabelli, L., Argentini, C., Ruiz, L., Margolles, A., van Sinderen, D., Turroni, F., Ventura, M. "**Genetic insights into the dark matter of the mammalian gut microbiota through targeted genome reconstruction**"
(2021) Environmental Microbiology, doi:10.1111/1462-2920.15559.

31) Mancabelli, L., Tarracchini, C., Milani, C.,Lugli, G.A., **Fontana, F.,** Turroni, F., van Sinderen, D., Ventura, M. "**Vaginotypes of the human vaginal microbioma**".
(2021) Environmental Microbiology, doi:10.1111/1462-2920.15441.

32) Tarracchini, C., Milani, C., Lugli, G.A., Mancabelli, L., **Fontana, F.,** Alessandri, G., Longhi, G., Anzalone, R., Viappiani, A., Turroni, F., van Sinderen, D., Ventura, M. "**Phylogenomic disentangling of the bifidobacterium longum subsp. infantis taxon**"
(2021) Microbial Genomics, doi:10.1099/MGEN.0.000609.

33) **Fontana, F.,** Alessandri, G., Lugli, G.A., 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**"
(2021) Microorganisms, doi:10.3390/microorganisms9010073.