UNIVERSITÁ DEGLI STUDI DI NAPOLI "FEDERICO II"

DIPARTIMENTO DI AGRARIA

DOTTORATO IN FOOD SCIENCE

XXXV CICLO

Identification of species and sub-species in food and human microbiomes

Tutor:
Ch.mo Prof. Edoardo Pasolli

Candidato:
Italia Elisa Mauriello

Coordinatore:
Ch.ma Prof.ssa Amalia Barone

2021-2022

Summary

The term microbiome refers to the whole community of living microorganisms in a sample along with their potential activities that might influence the metabolic capabilities and functioning of such a micro-environment. The study of the human microbiome has assumed a central role in the scientific community due to its fundamental role in health and disease, especially in the last few years thanks to the advancement of sequencing technologies and the dramatic decrease of sequencing cost for exploring complex microbial communities. This is generating a large amount of data that needs proper computational tools and methodologies to be analysed.

The overall aim of my doctoral research activity is to develop and apply tools and analyses for the identification of species and sub-species in food and human microbiomes. This is accomplished by considering large-scale approaches built on shotgun metagenomics data. In chapter 2, it has been developed and validated a methodology based on a clustering approach for sub-species identification from genomes. The main idea is to view clustering as a supervised classification problem, in which we must estimate the true number of clusters (i.e., sub-species in our case). We tested prediction strength which is the methodology mostly used in the microbiome field for this purpose, and proposed an alternative solution based on clusterwise cluster stability assessment by resampling which exhibited higher accuracies and reduced computational times. Such methodology has been validated on synthetic data in which we tried to estimate the right number of clusters/sub-species by changing different variables such as number of clusters and distances among clusters. The methodology has been also applied in real scenarios by considering two microbial families of great relevance in the food science field as lactic acid bacteria (LAB) and Bifidobacteriaceae.

In chapter 3, we performed large-scale genome-wide analysis for LAB species by considering microbiomes from both food sources and human body sites. We investigated the prevalence and diversity of LAB species in the human microbiome and their overlap with species and strains found in food. Quantitative taxonomic profiling was applied on the entire set of genomes, comprising food metagenomes coming from previously published studies and others newly sequenced in this study, and publicly available metagenomes corresponding to the human microbiome. The metagenome assembled genomes (MAGs) coming from human and food sources were integrated with publicly available genomes and subjected to extensive comparative genomic analyses.

A similar effort was devoted to the characterization of species and sub-species belonging to the *Bifidobacteriaceae* family in chapter 4. We estimated prevalence and abundance of *Bifidobacteriaceae* in the human microbiome, followed by sub-species identification and characterization. In this way, we tried to close the gap between human microbiomes and *Bifidobacteriaceae* strains that we commonly find in probiotic supplements. Also in this case we considered both MAGs and reference genomes, clustered them for species and sub-species identification, and conducted comparative genomic analyses. Finally, we built machine learning based predictive models to evaluate to which extent phenotype characteristics could be estimated by occurrence of *Bifidobacteriaceae* species.

An overview of the performed research activity along with future directions are described in the final chapter 5.

Table of contents

1 Introduction

1.1 An introduction to the human microbiome

The term microbiome refers to a characteristic microbial community occupying a reasonably well-defined habitat which has distinct physio-chemical properties. The term not only refers to the microorganisms involved but also encompass their theater of activity. The latter involves the whole spectrum of molecules produced by the microorganisms, including their structural elements (nucleic acids, proteins, lipids, polysaccharides), metabolites (signaling molecules, toxins, organic, and inorganic molecules), and molecules produced by coexisting hosts and structured by the surrounding environmental conditions. Therefore, all mobile genetic elements, such as phages, viruses and extracellular DNA, should be included in the term microbiome [1].

The human body consists of 10-100 trillion of symbiotic microbial cells and the human microbiome refers to the genes these cells harbor [2]. The study of the human microbiome has assumed a central role in the scientific community due to its fundamental role in health and disease [2–5], especially in the last few years thanks to the advancement of sequencing technologies and the dramatic decrease of sequencing cost for exploring complex microbial communities.

The microorganisms colonize various sites on and in the human body, where they adapt to specific features of each niche (e.g., skin, gut, mouth, and vagina). A great variation in both composition and function is observed when comparing one body niche to another [6]. The majority of microorganisms found in the body live in the human gut.

Moreover, the human microbiome has the potential to uniquely identify individuals, much like a fingerprint. Personal microbiomes contain enough distinguishing features to identify an individual over time. While individuals from the same human population usually contain similar species, different people typically carry person-specific strains [7].

1.2 Not only human, the food microbiome

The microbial community found in certain types of food represents the food microbiome. The relationship between foods and their microbiome is fundamental to their quality and safety. Beneficial microbial communities can be responsible for rheological and organoleptic traits of

fermented foods. However, undesirable microbes may also be present, and their development may affect the quality of food, leading to spoilage or other food safety issues [8]. Metataxonomics and metagenomics are currently the gold standard methodologies to explore the full potential of metagenomes in the food industry [9]. Metagenomics and metataxonomics display a different, although complementary, perspective. While metataxonomics does not provide information about the functional and metabolic features of the microorganisms and it is limited to depicting a profile of the members of the community, metagenomics exploits the information present in the whole genetic content of the community (the metagenome), usually by directly sequencing the total DNA pool of the microbial population, avoiding the bias introduced by the amplification of specific DNA fragments performed in metataxonomics approaches. The sequencing of all microbial DNA present in a sample has been defined as shotgun metagenomics, which currently is the gold standard to analyses complex microbial communities [10]. Shotgun metagenomics has been extensively used to depict the microbiomes of different environments, including those associated with foods. The food metagenome has been studied through shotgun sequencing in both non-fermented foods, such as milk and honey [11,12], and fermented foods, the latter being the ones that have received the most attention because their microbial load is normally high. Among fermented foods, shotgun sequencing methods have been applied in the cheese industry to assess the functional features of the microbiota of cow's milk artisanal cheeses from Northwestern Argentina, which has contributed to the isolation of bacteriocin-producing bacteria against Listeria monocytogenes [13]. The most comprehensive metagenomic analysis of different cheese types showed the usefulness of shotgun sequencing to link different bacterial functionalities, such as the synthesis of volatile compounds during ripening or bacteriocin-production, with genes or bacteria present in the cheese microbiota, providing a tool to improve cheese production processes [14]. In addition, metagenomes of fermented-meat and meat-processing industries have also been investigated. The potential functions associated with meat fermentation processes have been studied in sausages, highlighting the key role of the starter cultures in the organoleptic properties of fermented products [15].

Overall, as demonstrated in previous examples, the knowledge generated through shotgun metagenomics on food and food processing environments can help towards the selection of starter and adjunct bacterial cultures capable of conferring desired quality attributes to the final product, either in terms of improved nutritional, functional, or organoleptic properties. But it can also help to improve its safety through selecting microorganisms capable of extending their

shelf-life and to guarantee the absence of spoilage or pathogenic bacteria in a range of food products.

Finally, it is important to note that there are still great challenges to be solved in the shotgun analysis field, including the difficulties of analysing foods and food-related environments with a low microbial load with the currently available methodologies, as well as the lack of specific bioinformatics pipelines adapted to the study of food microbiomes. Therefore, there is a need to fine tune current shotgun approaches to fully explore the potential of these applications and implement these new methodologies in the food industry, which will undoubtedly contribute to the increase of quality and safety of food [9].

1.3 Identification of species and sub-species from metagenomic data

Microbiome research has been strongly driven by advances in DNA sequencing technologies, often referred to as next-generation sequencing, NGS [16]. With the advent of DNA sequencing and high-throughput technologies applied in all fields of biological sciences, we are able to generate billions of data points, which can be used for an in-depth characterization of the structure, function, inter-action, and complexity of microbial ecosystems. This has stimulated the development of sophisticated bioinformatics tools to analyze the massive amounts of data generated.

The two main approaches for analyzing the microbiome are 16S ribosomal RNA (rRNA) gene amplicons and shotgun metagenomics. 16S sequencing is used to identify and classify microbes by selectively amplifying and sequencing the hypervariable regions of the 16S rRNA gene. On the other hand, shotgun sequencing is less subject to amplification bias than 16S sequencing because it does not rely on targeted primers to amplify a marker gene. Shotgun metagenomics is usually more expensive but offers increased resolution, enabling a more specific taxonomic and functional classification of sequences as well as the discovery of new bacterial genes and genomes [17]. Importantly, shotgun metagenomics allows the simultaneous study of archaea, viruses, virophages, and eukaryotes [18,19].

From shotgun sequencing, it is now possible to construct metagenome-assembled genomes (MAGs). A MAG refers to a group of scaffolds with similar characteristics from a metagenome assembly that represent the microbial genome [20]. In this approach, sequencing reads are

assembled into scaffolds and then the scaffolds are grouped into candidate MAGs based on tetranucleotide frequencies (TNFs), abundances, complimentary marker genes [21], taxonomic alignments [22] and codon usage [23]. The MAGs with high completeness and low levels of contamination are then used for further taxonomic annotation and large-scale analyses. An established approach to detect the many species belonging to the human microbiome is to group the MAGs in species-level genome bins (SGBs), as described in [24].

Despite often being the highest resolution taxonomic category considered in microbiome surveys, species can contain extreme phenotypic variability [25]. Diversity within bacterial species is the result of continuous processes of variation generation due to mutations and gene flow mechanisms. Mutations are changes in DNA sequence and arise continuously in the genome owing to errors in the DNA replication process, damages caused by mutagens, or errors in the DNA repair and recombination mechanisms [26]. Mutations that arise in one genome can be passed vertically to descendants or horizontally to neighbouring cells. Usually, mutations increase the amount of variation within a species. Gene flow is the transfer of genetic variation from one population to another that can cause additions and rearrangements of genomic regions [27]. In terms of impact on within-species variation, the most important factor of the transfer is not the mechanism but rather whether or not the genetic material being transferred is novel to the recipient species. Natural selection and genetic drift determine the fate of within-species variability introduced through mutation and gene flow. Genetic drift randomly eliminates genetic variations within a population, whereas natural selection maintains or eliminates variations that respectively confer a fitness advantage or disadvantage. Diversity within species is generated, maintained and purged to different extents, such that some species are highly heterogeneous whereas others are tightly cohesive.

These features of within-species variation depend on the populations observed. At one extreme, species can be monotypic; that is, they have a uniform distribution of genetic similarities across their entire population. Monotypic species with low diversity are more likely to be specialists, with narrow geographic distributions or host ranges [28,29]. At the other extreme, species with subspecies (polytypic) and high diversity are more likely to be free-living generalists with multiple adaptations to distinct and fluctuating environments as well as broad geographic ranges or many partially overlapping niches [29,30].

Within-species genetic variation can be measured in many ways, with some common metrics being overall genome similarity, the number of shared and unique genes, and/or the number and nature of SNVs (single nucleotide variants). The overall similarity between genomes

belonging to the same species at higher resolution levels can be assessed from metagenomic data either directly from reads and reference genomes [31] or through comparison of MAGs [32]. Reference genome-based approaches can be limited by the low availability of appropriate reference genomes. Instead, large sets of MAGs are now available, and methods to calculate ANI (average nucleotide identity) have improved in efficiency [33]. However, calculating ANI for large genomic cohorts remains computationally challenging [32] and using MAGs can introduce inaccuracies owing to data quality limitations and incompleteness. The range and distribution of ANI values within species vary by taxon and population [32] and therefore, in contrast to species boundaries, within-species variants do not seem to display a universal threshold that would categorize them into groups. Further, genetic differences that are coded by a small number of nucleotides relative to the size of the genome, and thus have a small impact on ANI, can have a very large impact on phenotype. In these cases, measures of gene content and SNVs can be more informative than ANI for defining biologically relevant within-species variants.

The range of genetic variation within species can be covered by three terms: genome, strain and subspecies (**Figure 1.1**). A species potentially contains multiple subspecies, a subspecies contains multiple strains and a strain contains multiple (non-identical) genomes. These genomes can be sequenced from cultured isolates (isolate genomes) or through assembly of a metagenomic sample (MAGs). The former usually represents a cultured isolate with little diversity, whereas the latter might represent a population containing considerable diversity [25]. A universally applicable, operational definition of strain with a strong biological basis has not been established and may not exist. In theory, genomes with a few SNV differences could be referred to as different strains. Subspecies are clusters of strains that are genetically or phenotypically distinct.

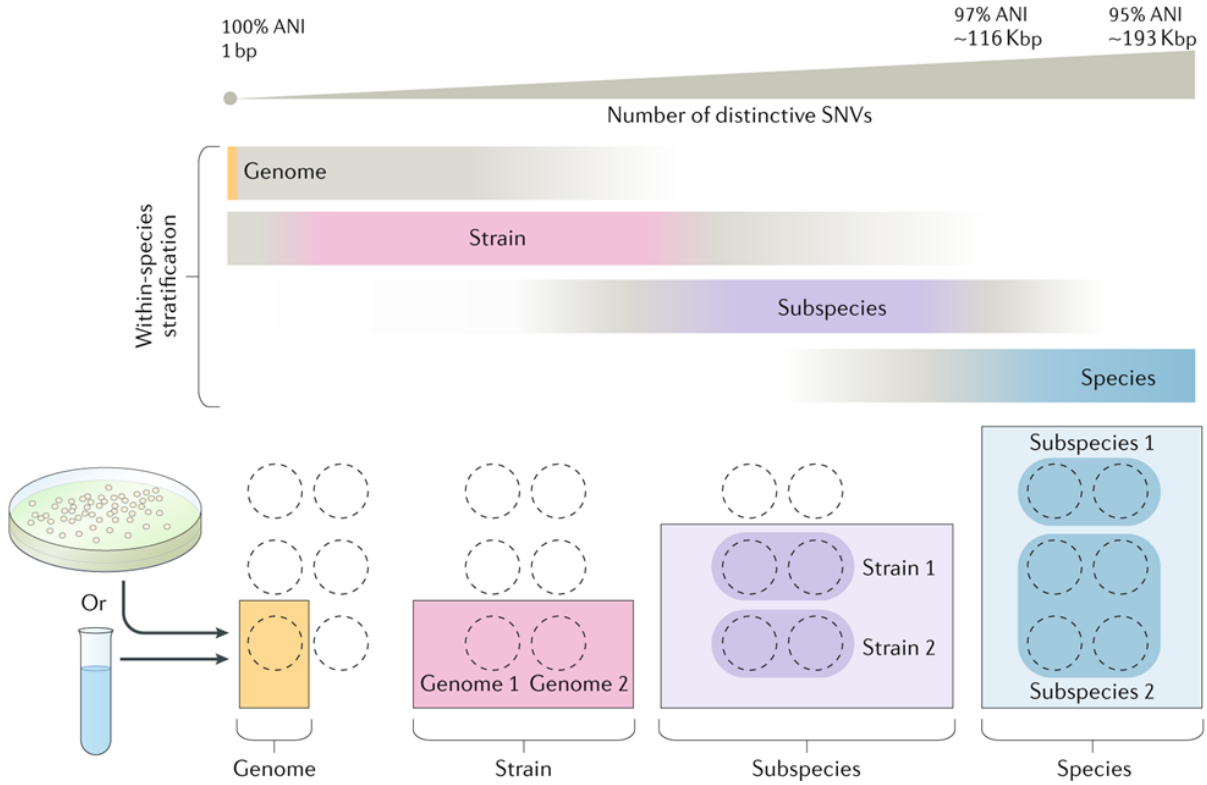


Figure 1.1. Within-species stratification
Terminology used to stratify variation within bacterial species, ranging from a single nucleotide variant (SNV) in the whole genome to the species-level threshold (95% ANI). [Figure is taken from [25]].

SGBs detection allows to characterize the human microbiome at most at species-level, but it is more challenging to characterize it at within-species-level by identifying sub-species within each specific SGB.

Along with several efforts performed in the literature to identify species in the microbiome from genomes and metagenomics data, a more open question is represented by the identification of sub-species and sub-clades. Metagenomic sequencing and technical advances have enabled culture-free, high-resolution strain and subspecies analyses at high throughput and in complex environments [25]. Overall, the identification of sub-species can be also relevant to perform more in-depth analyses into some specific species of interest.

1.4 Summary of my doctoral research activity

The overall aim of my doctoral research activity is to develop and apply tools and analyses for the identification of species and sub-species in food and human microbiomes. This is

accomplished by considering large-scale approaches built on shotgun metagenomics data. In chapter 2, it has been developed and validated a methodology based on a clustering approach for sub-species identification from genomes. The main idea is to view clustering as a supervised classification problem, in which we must estimate the true number of clusters (i.e., sub-species in our case). We tested prediction strength [34] which is the methodology mostly used in the microbiome field for this purpose, and proposed an alternative solution based on clusterwise cluster stability assessment by resampling which exhibited higher accuracies and reduced computational times. Such methodology has been validated on synthetic data in which we tried to estimate the right number of clusters/sub-species by changing different variables such as number of classes and distances among classes. The methodology has been also applied in real scenarios by considering two microbial families of great relevance in the food science field as lactic acid bacteria (LAB) and Bifidobacteriaceae.

In chapter 3, we performed large-scale genome-wide analysis for LAB species by considering microbiomes from both food sources and human body sites. We investigated the prevalence and diversity of LAB species in the human microbiome and their overlap with species and strains found in food. Quantitative taxonomic profiling was applied on the entire set of genomes, comprising food metagenomes coming from previously published studies and others newly sequenced in this study, and publicly available metagenomes corresponding to the human microbiome. The MAGs coming from human and food sources were integrated with publicly available genomes and subjected to extensive comparative genomic analyses.

A similar effort was devoted to the characterization of species and sub-species belonging to the *Bifidobacteriaceae* family in chapter 4. We estimated prevalence and abundance of *Bifidobacteriaceae* in the human microbiome, followed by sub-species identification and characterization. In this way, we tried to close the gap between human microbiomes and *Bifidobacteriaceae* strains that we commonly find in probiotic supplements. Also in this case we considered both MAGs and reference genomes, clustered them for species and sub-species identification, and conducted comparative genomic analyses. Finally, we built machine learning based predictive models to evaluate to which extent phenotype characteristics could be estimated by occurrence of *Bifidobacteriaceae* species.

An overview of the performed research activity along with future directions are described in the final chapter 5.

2 An automatic clustering-based approach to identify sub-species from genomes

2.1 Introduction and scientific rationale

2.1.1 Strategies to detect SGBs from genomes and MAGs

The human microbiome plays a role in health, but its full diversity remains uncharacterized. An established approach to characterize the many unidentified species belonging to the human microbiome is to reconstruct the genomes coming from the human microbiome (MAGs) and group them in species-level genome bins (SGBs), as described in [24]. In particular, SGBs were obtained by considering both reconstructed and isolate genomes by clustering based on whole-genome nucleotide similarity estimation of all of them. The cutoff on the hierarchical clustering was tuned based on the intra- and inter-species diversity of the confidently taxonomically labeled subset of the considered reference genomes.

In general, the thus obtained SGBs were further divided into known SGBs (kSGBs) that contain at least one reference genome and unknown SGBs (uSGBs) without any reference genomes. The kSGBs were then taxonomically labeled with the species label (if available) of the reference genomes present in the bin, whereas uSGBs were assigned to the phylum of their closest reference genome, and to a genus-level and family-level annotation when possible.

2.1.2 Identification of sub-species and sub-clades within each specific SGB

SGBs detection allows to characterize the human microbiome at most at species-level, but it is more challenging to characterize it at strain-level by identifying sub-species within each specific SGB. In general, the assessment of subspecies is essential since differences can arise within so defined species. The SGBs do not always have an assigned taxonomy, as described above, and therefore they are unknown in literature.

Therefore, the estimation of subspecies requires computational techniques based on clustering methods, in which the obtained clusters are equivalent to the subspecies for each SGB. The key

idea is to view clustering as a supervised classification problem, in which we must estimate the true number of clusters.

2.1.3 Strategies proposed in the literature for the detection of subspecies

For that purpose, in the literature there are some studies in which the identification of subspecies was performed by using the prediction strength metric [34].

For example, Costea et al. [35] conducted a large-scale survey of population structure in prevalent human gut microbial species, sampled from their natural environment, with a culture-independent metagenomic approach. They delineated population structure corresponding to subspecies. In particular, they used metaSNV [36] to compute a matrix of genomic allele distances between all samples that allowed variant calling for any given species and identified clusters by applying a very stringent cutoff for separation: they used the prediction strength to determine the support of the PAM clustering for each number of clusters k; the highest number of clusters that have a prediction strength above 0.8 was considered to be the number of subspecies, as recommended by Tibshirani and Walther for determining high-quality clusters [34]. With this operational definition, they found between two and four subspecies in 44 of the 71 considered species (accounting for 72% (SD = 15%) of the assigned relative abundance per sample), totaling 112 subspecies (**Figure 2.1**). They also assessed the global geographic distribution of each subspecies. While many subspecies appeared to be distributed without any recognizable geographic pattern, some did show striking regional enrichments.

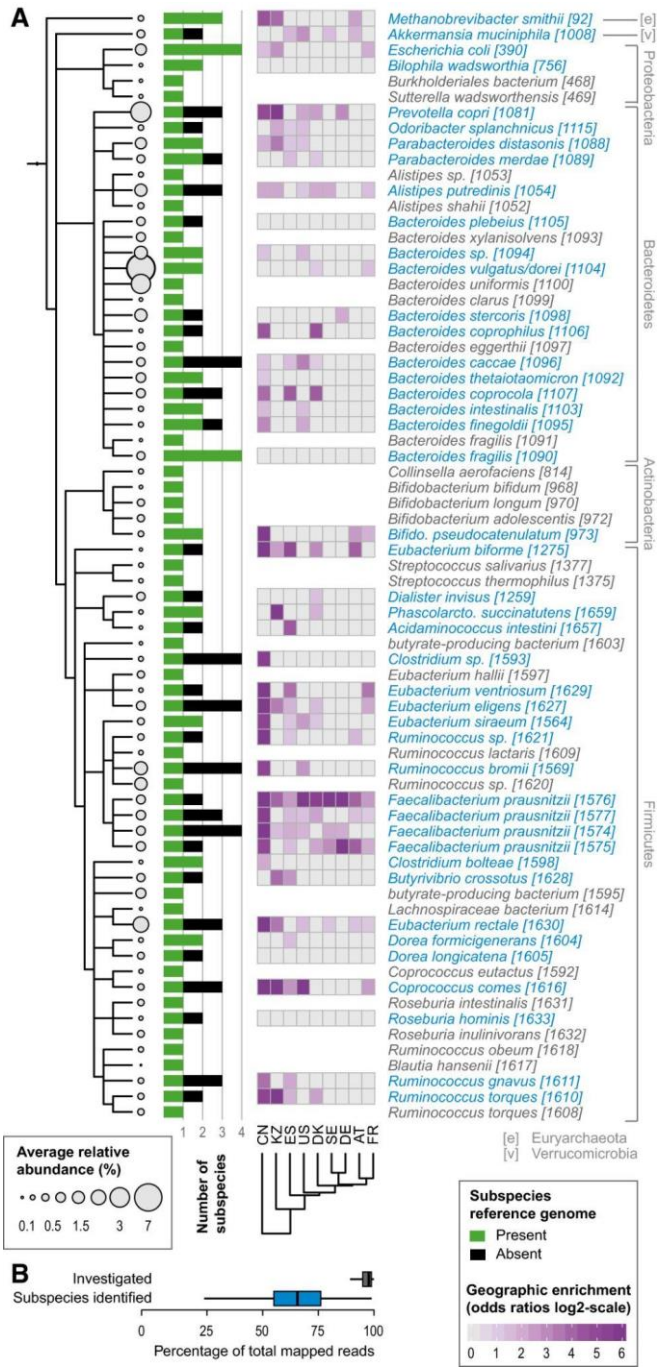


Figure 2.1. Identification and prevalence of human gut microbial subspecies

A,**B)** Human gut microbial species explored for the existence of subspecies show wide phylogenetic spread according to NCBI taxonomy **(A)** and include *Methanobrevibacter smithii*, the main archaeal member of the human gut microbiome, as well as representatives of all abundant phyla. Species names are according to NCBI taxonomy, with species cluster (specI) identifiers according to Mende et al. [37], which splits some named species into multiple specI clusters. Of 71 investigated species, 44 stratify into subspecies (highlighted in blue). Each species' average abundance across 2,144 human gut metagenomes is proportional to the size of the circles on the cladogram. Bars represent the number of subspecies identified in each, with "1" indicating no subdivision. The black portion of the bar corresponds to subspecies for which no representative genome sequence is available from NCBI. Geographic enrichments of subspecies are displayed as a heat map (showing only significant enrichment, FDR-corrected Fisher test P-value < 0.05, per country as maximum log-odds ratio across

conspecific subspecies). Subspecies with a restricted geographic range are predominantly found in the Chinese and Kazakh populations. The 71 investigated species captured an average of 95.5% of sequencing reads that were assigned to any reference genome. The subset of 44 species with identified subspecies accounted for the majority of this abundance **(B)**. [Figure with associated caption is taken from [35]].

Moreover, they investigated whether subspecies occurrences are also restricted in individuals and how stable an individual's gut subspecies composition is over time. They observed a general dominance or exclusivity of a single subspecies that highlights gut microbial individuality. Overall, that persists over time under normal conditions.

The identification of subspecies with prediction strength was applied in a study that investigated the global distribution and population structure of *Prevotella copri* [38], a common human gut microbe that has been both positively and negatively associated with host health. In this study it was assessed if the presence of *P. copri* is associated with different human diseases. This analysis revealed that *P. copri* is not a monotypic species but is composed of four distinct clades. In a meta-analysis of available disease phenotypes, the authors found no strong evidence that any of the four clades were associated with a disease. Specifically, to investigate the association of the *P. copri* complex with different diseases, they analyzed the prevalence and abundance of the four clades for each cohort where the study design included both case and controls. At the clade level, there is no clear evidence to suggest P. copri is associated with the etiology of the considered diseases. Extending the analysis further to consider sub-clades also did not reveal any statistically significant associations with disease.

The implementation of prediction strength in a large-scale analysis of the genomes belonging to one of the most prevalent human gut bacteria, i.e. *Eubacterium rectale* [39], highlights the existence of four subspecies (prediction strength consistently over 0.8 for k = 4), one of which was not observed before that study. Three of these four subspecies are large and well-defined monophyletic subtrees in the phylogeny, and only a minority of strains of the four *E. rectale* subspecies showed very strong geographic enrichment. In particular, the three most represented subspecies predominantly comprised strains from Europe and Asia. The fourth and previously unobserved subspecies, included strains derived mostly from sub-Saharan African countries but also contains strains from Peru and Indonesia. Moreover, to assess the possibility of inter-individual *E. rectale* strain transmission in human populations, they further analyzed metagenomic data from mother-infant pairs in multiple cohorts and found evidence of vertical transmission. Their analyses suggest that *E. rectale* is specific to humans and that it can be

transmitted within populations. They found a statistically significant correlation (p value 0.041) between pairwise geographic and median genetic distances of subspecies that is confirmed when directly considering pairwise distances between samples (p value <1e−16), suggesting that *E. rectale* genetic stratification could have been to some extent shaped by physical isolation of strains over time.

2.1.4 An alternative solution to identify subspecies

We tested prediction strength but we propose a new solution to improve the accuracy and reduce computational time. The proposed methodology is based on clusterboot (Clusterwise cluster stability assessment by resampling) [40].

We performed both prediction strength and clusterboot on simulated data and the obtained results highlighted that our alternative strategy based on clusterboot works better in terms of assessment of clusters and computational time.

We then applied our strategy in a real context that is the identification of subspecies (clusters) in large-scale analysis of bacterial species. In the further chapter 4 we described an analysis conducted on the *Bifidobacteriaceae* family delineating novel subspecies in its prevalent species.

2.2 Materials and methods

2.2.1 Definition of SGBs starting from genomes/MAGs through a clustering approach

We hypothesize having a set of genomes that can be obtained from both isolate and metagenomic sources and would like to group them in species-level genome bins (SGBs) (**Figure 2.2A**). The genomes were clustered based on whole-genome nucleotide similarity estimation using Mash [33]. Mash enables the comparison and clustering of whole genomes and metagenomes on a massive scale; it reduces large sequences and sequence sets to small, representative sketches, from which global mutation distances can be rapidly estimated. Methods based on string matching can produce very accurate estimates of mutation distance, but must process the entire sequence with each comparison, which is not feasible for all-pairs

comparisons. In contrast, the Mash distance can be quickly computed from the size-reduced sketches alone, yet produces a result that strongly correlates with alignment-based measures. Thus, Mash combines the high specificity of matching-based approaches with the dimensionality reduction of statistical approaches, enabling accurate all-pairs comparisons between many large genomes and metagenomes.

The computed distances matrix by Mash is the input for hierarchical clustering, a method of cluster analysis in which the data are grouped into a tree of clusters. The cutoff on hierarchical clustering was tuned based on intra- and inter-species diversity of the confidently taxonomically labeled subset of the reference genomes resulting in SGBs spanning ~5% genetic diversity, as independently proposed elsewhere [32].

2.2.2 Estimation of the number of sub-species in each specific SGB

We would like to estimate the number of subspecies in each specific SGBs (**Figure 2.2B**). For this purpose, we consider the set of genomes falling in the considered SGB and estimate their pairwise average nucleotide identities (ANIs). This was performed through FastANI [32], which provides more accurate results than the more approximate distances estimated by Mash. ANI is the estimation of the genetic relatedness between two genomes; it represents the average nucleotide identity of all orthologous genes shared between any two genomes and offers robust resolution between strains of the same or closely related species (i.e., showing 80–100% ANI). The algorithm FastANI alleviates the computational bottleneck in ANI computation using alignment-free approximate sequence mapping. FastANI is accurate for both finished and draft genomes, and is up to three orders of magnitude faster compared to alignment-based approaches.

Then we assessed the presence of subspecies through visual inspection in multidimensional scaling (MDS) plot using the Partitioning Around Medoid (PAM) algorithm as clustering method. In particular, we used the "cmdscale" function available in the R stats package to compute the MDS. The R function "cmdscale" stands for Classical (Metric) Multidimensional Scaling that is also known as principal coordinates analysis [41]. Multidimensional scaling takes a set of dissimilarities and returns a set of points such that the distances between the points are approximately equal to the dissimilarities. A set of euclidean distances on n points can be represented exactly in at most $n - 1$ dimension; "cmdscale" returns the best-fitting k-dimensional representation, where k may be less than the maximum dimension of the space

which the data are to be represented in, that is must be in $[1, 2, \ldots, n-1]$. The representation is only determined up to location, rotation and reflections. The configuration returned is given in principal-component axes, so the reflection chosen may differ between R platforms. The PAM algorithm is intended to find a sequence of objects called medoids that are centrally located in clusters; the goal of the PAM algorithm is to minimize the average dissimilarity of objects to their closest selected object. Equivalently, it can minimize the sum of the dissimilarities between an object and their closest selected object.

The number of sub-species was estimated by identifying the optimal number of clusters for each SGB. We determined the optimal number of clusters by considering and comparing two different methodological approaches: prediction strength and clusterboot. Both methodologies rely on the main idea to view clustering as a supervised classification problem in which we have to estimate the true class labels.

We finally validated our results using a phylogenetic tree, which is a diagrammatic representation of the evolutionary relationships among various taxa. It is a branching diagram composed of nodes and branches. The branching pattern of a tree is called the topology of the tree. The nodes represent taxonomic units, such as species (or higher taxa), populations, genes, or proteins. A branch is called an edge, and represents the time estimate of the evolutionary relationships among the taxonomic units. One branch can connect only two nodes. In a phylogenetic tree, the terminal nodes represent the operational taxonomic units (OTUs) or leaves. The OTUs are the actual objects, such as the species, populations, or gene or protein sequences, being compared, whereas the internal nodes represent hypothetical taxonomic units (HTUs). An HTU is an inferred unit and it represents the last common ancestor (LCA) to the nodes arising from this point. We verified that the genomes belonging to the same subspecies were close in the considered SGB's phylogenetic tree.

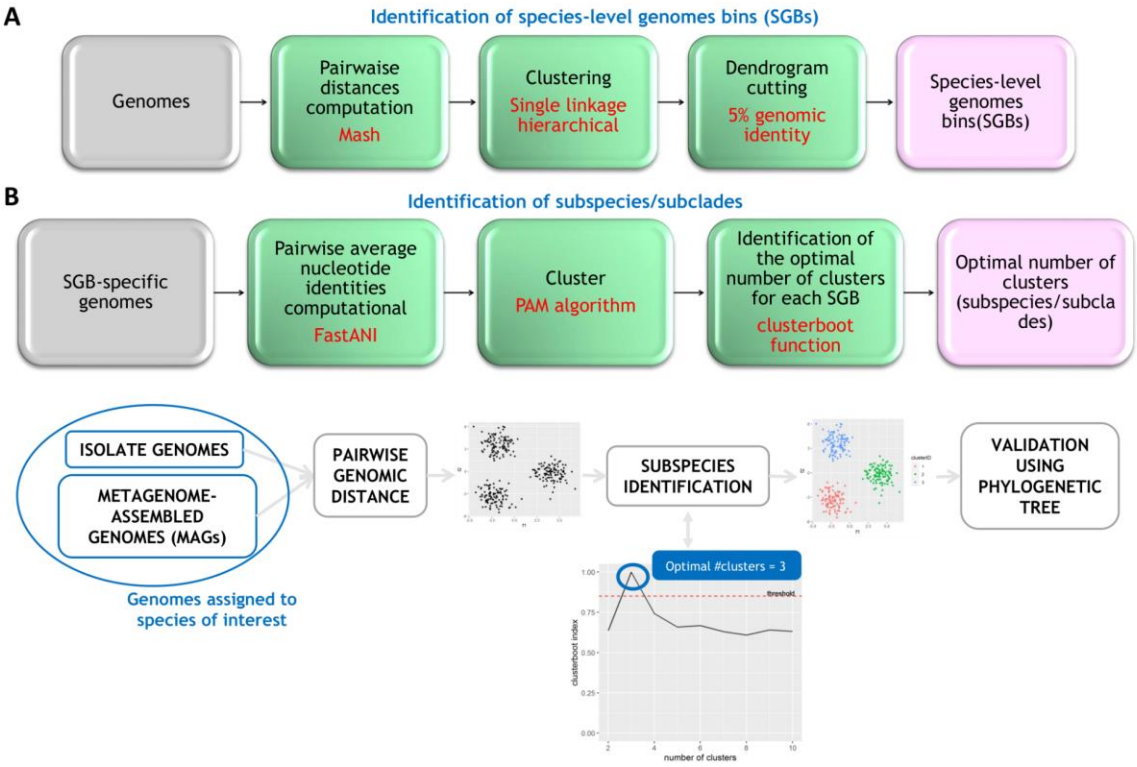


Figure 2.2. SGBs and subspecies identification workflow
A) The identification of SGBs was performed by applying hierarchical clustering (threshold of ~5% genetic diversity) on pairwise distance matrix computed using Mash on the whole set of considered genomes. B) From the SGB-specific genomes we computed the pairwise genetic distance using FastANI and estimated the optimal number of subspecies through "clusterboot" R function. We assessed the presence of subspecies through visual inspection in MDS plot using PAM as clustering method and we finally validated the results using phylogenetic trees.

2.2.3 Methodological description of prediction strength

The prediction strength measure assesses how many groups can be predicted from the data, and how well. To describe how prediction strength works we report here the analytical details copied from the original work [34].

The authors considered a training data $X_{tr} = \{x_{ij}\}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$, consist of p features measured on n independent observations. Let $d_{ii'}$ denote the distance between observations i and i'. The most common choice for $d_{ii'}$ is the squared Euclidean distance $\Sigma_j (x_{ij} - x_{i'j})^2$. Suppose we cluster the data into k clusters. For example, we might use k-means clustering based on Euclidean distance, or hierarchical clustering. Denote this clustering operation by $C(X_{tr}, k)$. Now when they apply this clustering operation to the training data,

each pair of observations either does or does not fall into the same cluster. To summarize this, let $D[C(\dots), X_{tr}]$ be an $n \times n$ matrix, with ii'th element $D[C(\dots), X_{tr}]_{ii'} = 1$ if observations i and i' fall into the same cluster, and zero otherwise. They call these entries "co-memberships". In general, the clustering $C(\dots)$need to be derived from X_{tr}.

For example, they can apply the k-means algorithm to some dataset Y, which will result in a partition of the observation space into k polygonal region of $C(Y, k)$. Their proposal for real data uses repeated cross-validation. To motivate this approach, consider the conceptually simpler scenario in which an independent test sample X_{te} of size m is available, drawn from the same population as the training set. As above, they can cluster X_{te} into k clusters via an operation $C(X_{te}, k)$, and summarize the cluster co-memberships via the $m \times m$ matrix $D[C(X_{te}, k), X_{te}]$. The main idea is to cluster the test data into k clusters, cluster the training data into k clusters, and then measure e how well the training set cluster centers predict co-memberships in the test set. For each pair of test observations that are assigned to the same test cluster, we determine whether they are also assigned to the same cluster based on the training centers. Here is the idea in detail. For a candidate number of clusters k, let $A_{k1}, A_{k2}, \dots A_{kk}$ be the indices of the test observations in test clusters $1, 2, \dots, k$. Let $n_{k1}, n, \dots n_{kk}$ be the number of observations in these clusters. They define the "prediction strength" of the clustering $C(., k)$ by

$$ps(k) = min_{1 \le j \le k} \frac{1}{n_{kj}(n_{kj}-1)} \sum_{i \neq i' \in A_{kj}} D[C(X_{tr}, k), X_{te}]_{ii'} \qquad (2.1)$$

For each test cluster, they compute the proportion of observation pairs in that cluster that are also assigned to the same cluster by the training set centroids. The prediction strength is the minimum of this quantity over the k test clusters. Here is the intuition behind this idea. If $k = k_0$, the true number of clusters, then the k training set clusters will be similar to the k test set clusters, and hence will predict them well. Thus $ps(k)$ will be high. Note that $ps(1) = 1$ in general, because both the training and test observations all fall into one cluster. However, when $k > k_0$, the extra training set and test set clusters will in general be different, and thus we expect $ps(k)$ to be much smaller. Using the minimum rather than the average expression (2.1) makes the procedure more sensitive in many-cluster situations.

Note that in general it would be difficult to compare the training and test clusterings by associating each of the k training clusters with one of the test clusters. By focusing only on the pairwise co-memberships in (2.1), they finesse this problem. The identity of the cluster

containing each observation is not considered: only its co-memberships in some clusters are used.

They choose the optimal number of clusters k' to be the largest k such that $ps(k)$ is above some threshold. Experiments show that a threshold in the range $0.8 \div 0.9$ works for well separated clusters. They think of k' as the largest number of clusters that can be reliably predicted in the dataset. Now in the absence of a test sample, the authors suggest instead use repeated r-fold cross-validation to estimate the prediction strength (2.1). The first $r-1$ folds represent the training sample, while the last fold is the test sample. Prediction strength for individual observations can also be defined. Specifically, the authors define the prediction strength for observation i as:

$$ps(i,k) = \frac{1}{\#A_k(i)} \cdot \sum_{i' \in A_k(i)} 1(D[C(X_{tr},k), X_{te}]_{ii'} = 1) \qquad (2.2)$$

where $A_k(i)$ are the observations indices i' such that $i \neq i'$ and $D[C(X_{tr},k), X_{te}]_{ii'} = 1$.

2.2.4 Description of the alternative strategy based on clusterboot

Many papers use stability or prediction strength measurements as a tool to estimate the true number of clusters. The alternative approach that we considered is based on clusterwise cluster stability assessment by resampling (clusterboot). We report here the methodological details as copied from the original paper [40]. The method has the following two important characteristics:

- It is applicable to very general clustering methods including methods based on (not necessarily metric) dissimilarity measures, non-partitioning methods and methods that include an estimator of the number of clusters, as well as conventional methods based on Euclidean data with a fixed number of clusters such as k-means. No particular cluster model is assumed.
- The approach is cluster-wise. The idea behind this is that many data sets contain meaningful clusters for which a certain cluster model is adequate, but they do not necessarily consist only of such clusters. Therefore, the result of a clustering method could find some important meaningful patterns in the data set, while other clusters in the same clustering can be spurious. The reason for this is not necessarily the choice of the wrong clustering method; it may well be that no single method delivers a satisfactory result for the whole data set.

The basic method is based on a non-parametric bootstrap. The author described the methodology as follows.

A sequence of mappings $E = (E_n)_{n \in N}$ is called a general clustering method, if E_n maps a set of entities $x_n = \{x_1, \ldots, x_n\}$ onto a collection of subset $\{C_1, \ldots, C_s\}$ of x_n. Note that it is assumed that entities with different indexes can be distinguished. This means that the elements of x_n are interpreted as data points and that $|x_n| = n$ is even if, for example, for $i \neq j$, $x_i \neq x_j$ in terms of their values. It is not assumed how the entities are defined. This could be, e.g., via a dissimilarity matrix or via p Euclidean variables. Most clustering methods generate disjoint clusterings, i.e., $C_i \cap C_j = \emptyset$ for $i \neq j \leq k$. A partition is defined by $\cup_{j=1}^{k} C_j = x_n$.

The methodology defined here does not necessarily assume that the clustering method is disjoint or a partition, but the interpretation of similarity values between clusters is easier for methods that do not generate a too rich clustering structure. For example, if the clustering method generates a full hierarchy, every subset containing only one point is always a cluster and these clusters will be perfectly stable, though totally meaningless. To assess the stability of a cluster of the initial clustering with respect to a new clustering, a similarity measure between clusters is needed. Because the measure should be applicable to general clustering methods (even methods that do not operate on the Euclidean space), it has to be based on set memberships. There exist many similarity measures between sets. He suggests the Jaccard coefficient, which originated in the analysis of species distribution data:

$$\gamma(C, D) = \frac{|C \cap D|}{|C \cup D|}, \qquad C, D \subseteq x_n \tag{2.3}$$

The Jaccard coefficient gives the proportion of points belonging to both sets of all the points involved in at least one of the sets, and it is therefore easily directly interpretable. It has several good properties, e.g., being independent of the number of points not belonging to any of the two sets. The Jaccard coefficient is used to compare cluster analysis methods theoretically, and the value $\frac{1}{2}$ was defined as a critical value for so-called "dissolution" of a cluster under addition of points to the data set. It can be shown that $\frac{1}{2}$ is the smallest value so that every cluster in a partition consisting of more than one cluster can be dissolved by a new partition, and it is also the smallest value so that whenever an initial cluster has s clusters and a new clustering has $r < s$ clusters, then at least $s - r$ clusters of the original clustering are dissolved.

The Jaccard coefficient has also been used in the context of cluster validation with resampling methods, though not for cluster-wise evaluation. The idea behind the use of the non-parametric

bootstrap for the assessment of cluster stability is the following: assume that there is an underlying mixture distribution $P = \sum_{i=1}^{s} \varepsilon_i P_i$ where P_i, $i = 1, \ldots, s$ are the distributions generating s "true" clusters, and ε_i is the probability that a point from P_i is drawn. For a given data set with n points, the "true" clustering would then consist of s clusters each of which contains exactly the points generated by P_i, $i = 1, \ldots, s$. When a data set generated from P is clustered, the found clusters differ from the "true" clusters, because the clustering method introduces a certain bias and variation. This can depend on the cluster P_i, for example, if two different clusters are weakly separated or if P_i deviates strongly from the cluster model assumed by the clustering method.

Bias and variation can be expressed by the maximum Jaccard coefficient between the set of all the points generated by P_i and actually obtained clustering. The bootstrap is usually used to give an idea of bias and variation caused by a certain statistical method, because in reality no true underlying distribution and no true clustering is known. The empirical distribution of the observed data set is then taken to simulate P. Points can be drawn from the data set and the originally found clusters can be treated as the "true" ones. The mean maximal Jaccard coefficient can be interpreted as indicating the stability of the original clusters. Given a number B of bootstrap replications and a cluster C from the original clustering $E_n(x)$, the scheme works as follows. Repeat for $i = 1, \ldots, B$:

1. Draw a bootstrap sample $x^i{}_n$ of n points with replacement from the original data set x_n.
2. Compute the clustering $E_n(x^i{}_n)$.
3. Let $x^i{}_* = x_n \cap x^i{}_n$ be the points of the original data set that are also in the bootstrap sample. Let $C^i{}_* = C \cap x^i{}_*$, $\Delta = E_n(x^i{}_n) \cap x^i{}_*$.
4. If $C^i{}_* \neq \emptyset$, compute the maximum Jaccard similarity between the induced cluster $C^i{}_*$ and the induced new clustering Δ on $x^i{}_*$: $\gamma_{C,i} = max_{D \in \Delta} \gamma(C^i{}_*, D)$ (i.e., D is the maximizer of $\gamma(C^i{}_*, D)$; else $\gamma_{C,i} = 0$).
5. This generate a sequence $\gamma_{C,i}$, $i = 1, \ldots, B$. The author suggests the mean

$$\gamma'_C = \frac{1}{B*} \sum_{i=1}^{B} \gamma_{C,i} \tag{2.4}$$

as stability measure ($B*$ being the number of bootstrap replications for which $C^i{}_* \neq \emptyset$ and is used here because in all other cases $\gamma_{C,i} = 0$).

Other summary statistics such as the median, a trimmed mean or the number of dissolutions ($\gamma_{C,i} \leq 0.5$) or good recoveries ($\gamma_{C,i} > 0.75$) can be used as well. Experience suggests that the mean is a good choice: in all examples in which were examined further statistics, It was not found any results that deviated strongly from what could be expected by looking at the mean alone. The value range and therefore also the size of possible outliers affecting the mean are restricted to $[0,1]$ and if moderate outliers occur, they may be treated as informative and need presumably not to be downweighted or trimmed. Generally, a valid, stable cluster should yield a mean Jaccard similarity value of 0.75 or more. Between 0.6 and 0.75, clusters may be considered as indicating patterns in the data, but which points exactly should belong to these clusters is highly doubtful. Below average Jaccard values of 0.6, clusters should not be trusted. Highly stable clusters should yield average Jaccard similarities of 0.85 and above [42].

2.2.5 Experimental setting

- For the assessment of these two methodological approaches, we generated simulated data with these characteristics:
    - The number of clusters N varied in the range $[2,3,4]$
    - The number of observations was set to $N * 100$ where N is the number of considered clusters.
    - We considered a number of features equal to the number of clusters. Values for each feature were generated by considering Gaussian random distributions having these characteristics: mean always equal to 0 apart for the i-th feature that was set to m (where i is the considered class number). In this way we guaranteed a certain distance among clusters; m was varied in the range $[0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5]$ to play with the distance among clusters. A value of m equal to 0 means that the distance between clusters is zero and therefore doesn't exist a real separation among clusters. In all cases standard deviation was equal to 1.
- We generated the input for the algorithms that provided us the estimation of optimal number of clusters computing Euclidean distance between observations.In a n-dimensional Euclidean space, let point p have Cartesian coordinates $(p_1, p_2, \ldots, p_n)$ and let point q have coordinates $(q_1, q_2, \ldots, q_n)$, the Euclidean distance between p and q is given by:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{2.5}$$

- We tested the more standard methodology based on prediction strength. We considered hierarchical clustering as a clustering method and we evaluated two different linkages, i.e. average and single. Single-linkage (nearest neighbor) is the shortest distance between a pair of observations in two clusters. It can sometimes produce clusters where observations in different clusters are closer together than to observations within their own clusters. These clusters can appear spread-out. Average-linkage is where the distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance. We computed the prediction strength value by varying the number of clusters between 2 and 10. We chose the estimated number of clusters as the maximum value above the threshold value; we considered a threshold to 0.8 as suggested in the original paper [34] and used in multiple microbiome papers. We therefore compared single and average linkages, and used the best found setting in the downstream comparisons.
- We compared the proposed solution based on the clusterboot methodology with respect to the prediction strength results. For clusterboot, we used the Clustering Large Applications (chapter 3 of [43]) algorithm for clustering ("clara"). Compared to other partitioning methods such as PAM, it can deal with much larger datasets. Internally, this is achieved by considering sub-datasets of fixed size such that the time and storage requirements become linear rather than quadratic. Each sub-dataset is partitioned into k clusters using the same algorithm as in PAM. Once k representative objects have been selected from the sub-dataset, each observation of the entire dataset is assigned to the nearest medoid. The mean (equivalent to the sum) of the dissimilarities of the observations to their closest medoid is used as a measure of the quality of the clustering. The sub-dataset for which the mean (or sum) is minimal, is retained. A further analysis is carried out on the final partition. Each sub-dataset is forced to contain the medoids obtained from the best sub-dataset until then. Randomly drawn observations are added to this set until the fixed size has been reached. Also in this case the number of clusters varied between 2 and 10. We considered two different thresholds based on the Jaccard similarities to estimate the optimal number of clusters: we wanted a mean value greater than 0.85 as suggested in the original paper [42], in addition to a minimum value greater than 0.8. This second criterion was implemented to improve the detection of the right number of clusters as we will show empirically later.
- We assessed performances of the methods in terms of:

- If the estimated number of clusters is equal to the true number of clusters. This is possible since we are considering synthetic data.
- Visual inspection through ordination analysis.
- Computational time. Empirical evaluation

2.3 Results

2.3.1 Average linkage works better than single linkage in prediction strength

We assessed the optimal number of clusters applying prediction strength measure testing two different linkages for hierarchical clustering, i.e. average and single. In single linkage the distance between two clusters is the minimum distance between members of the two clusters, whereas in the average linkage the distance between two clusters is the average of all distances between members of the two clusters. The **Table 2.1** shows that the prediction of number of clusters was always wrong (except one case) using single linkage.

Table 2.1. Evaluation of prediction strength in average linkage and single linkage
We computed prediction strength measure ("prediction.strength" R function) on simulated data as vary N (number of clusters) and m (value of mean that computes the distance among clusters). We report the number of estimated clusters and the error (no = number of estimated clusters is equal to real number of clusters, yes = otherwise) in tested linkages of hierarchical clustering, i.e. average and single.

		number of estimated clusters		**error**	
mean among clusters	**#clusters**	**average**	**single**	**average**	**single**
0	2	1	5	no	yes
0	3	1	6	no	yes
0	4	1	5	no	yes
0.5	2	1	4	yes	yes
0.5	3	1	7	yes	yes
0.5	4	1	4	yes	no
1	2	1	6	yes	yes
1	3	1	6	yes	yes
1	4	1	5	yes	yes
1.5	2	1	6	yes	yes

1.5	3	1	6	yes	yes
1.5	4	1	5	yes	yes
2	2	1	5	yes	yes
2	3	1	10	yes	yes
2	4	1	7	yes	yes
2.5	2	3	4	yes	yes
2.5	3	1	10	yes	yes
2.5	4	1	10	yes	yes
3	2	2	6	no	yes
3	3	1	8	yes	yes
3	4	1	9	yes	yes
3.5	2	2	4	no	yes
3.5	3	4	4	yes	yes
3.5	4	1	8	yes	yes
4	2	2	4	no	yes
4	3	4	10	yes	yes
4	4	4	2	no	yes
4.5	2	2	5	no	yes
4.5	3	4	10	yes	yes
4.5	4	4	1	no	yes
5	2	2	7	no	yes
5	3	4	10	yes	yes
5	4	4	10	no	yes

In **Figure2.3** we reported some examples to underline how average linkage works better than single. We picked up different configurations changing the number of clusters from 2 to 4 and the mean value equal to 4.5, 0 and 5 respectively. In all considered cases the estimation of the optimal number of clusters using average linkage works well. In **Figure2.3A** (N = 2, m = 4.5) we observed that the optimal number of clusters estimated with prediction strength in single

linkage is 5, even if the value is weakly above the threshold. In this case, the clustering consisted of 2 more populous clusters and three others very smaller ,as shown in the multidimensional scaling scatter plot. The second case (N = 3, m = 0) is the configuration without clusters, but single linkage results showed a clustering (**Figure2.3B**). In the last configuration (N = 4, m = 5) we should have 4 clusters well separated from each other due to the mean value being quite large. Prediction strength average linkage result was consistent, reaching a peak greater than threshold for 4 clusters. On the other hand, prediction strength single linkage result was unclear. The optimal number of clusters was 10 because is the maximum value tested (**Figure2.3C**).

In particular, we found, in all cases, that a single linkage as a metric to calculate clusters leads to overestimation of the optimal number of clusters; this is always true because it is due to the calculation of the linkage between two clusters using single linkage as measure. Therefore, we considered the results from average linkage to compare the examined metrics in the next analysis.

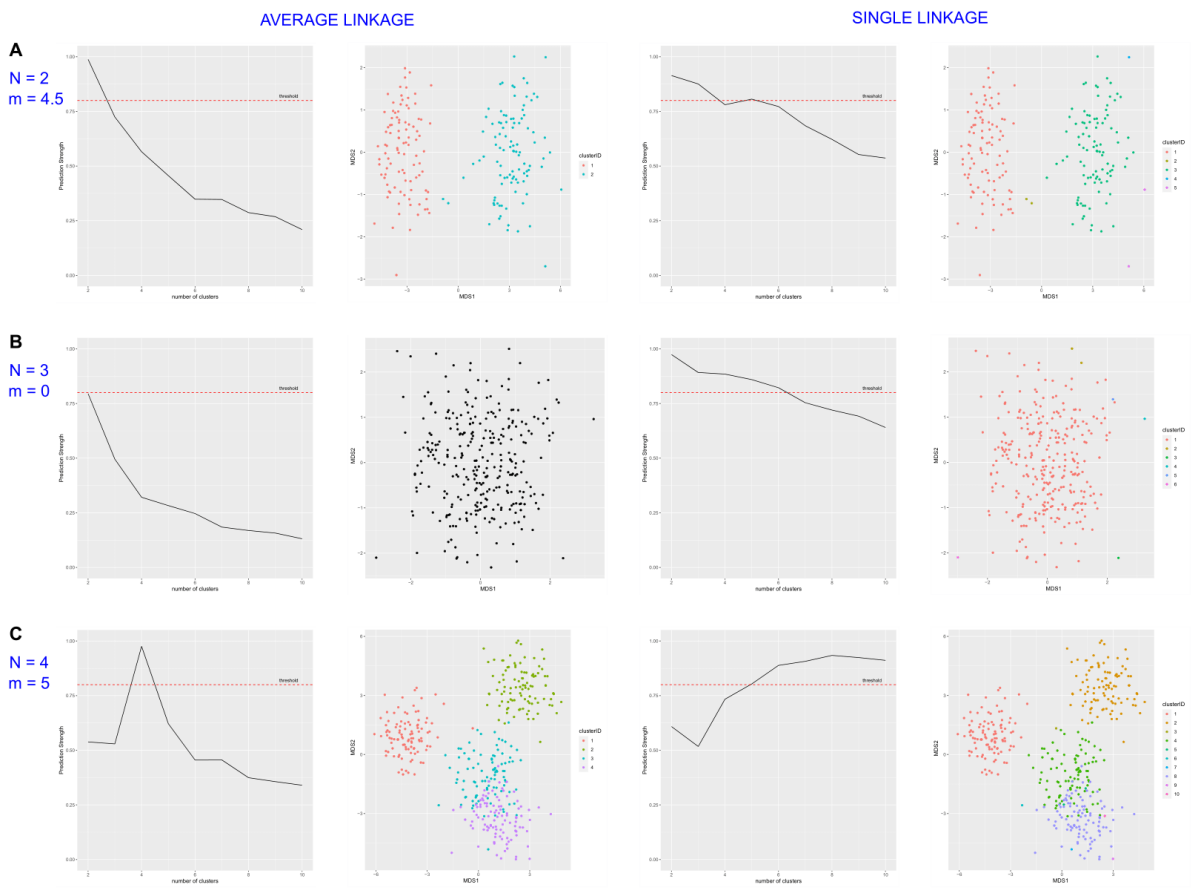


Figure 2.3. Prediction strength evaluation: Average linkage compared with single linkage
We report prediction strength results computed with "prediction.strength" R function and the corresponding MDS obtained with "cmdscale" R function with colored clusters, if any, for both linkages evaluated in hierarchical clustering: average linkage (on the left) and single linkage (on the right). We show different cases as vary N (number of clusters) and m (value of mean that computes the distance among clusters): A) N = 2 and m = 4.5; B) N = 3 and m = 0; C) N = 4 and m = 5.

2.3.2 Clusterboot requires two thresholds to estimate the cluster number

The configuration to estimate the optimal number of clusters with methodology based on clusterboot requires thresholding the Jaccard similarities. We used two criteria: once reported in the literature, i.e. the mean value of Jaccard similarities has to be greater than 0.85 [42], and the other concerning the minimum that has to be greater than 0.8. This other adopted criterion is useful to increase the robustness of the result avoiding possible mistakes in the assessment of the optimal number of clusters. In general, without this further assumption, we might overestimate the number of clusters.

As an example we reported in **Figure 2.4** the case of generated simulated data with 4 clusters and a mean value equal to 3. The optimal number of clusters estimated with clusterboot

considering only the threshold on mean value reported in literature was 5 instead of 4. If we considered the additional threshold on minimum we achieved the right result.

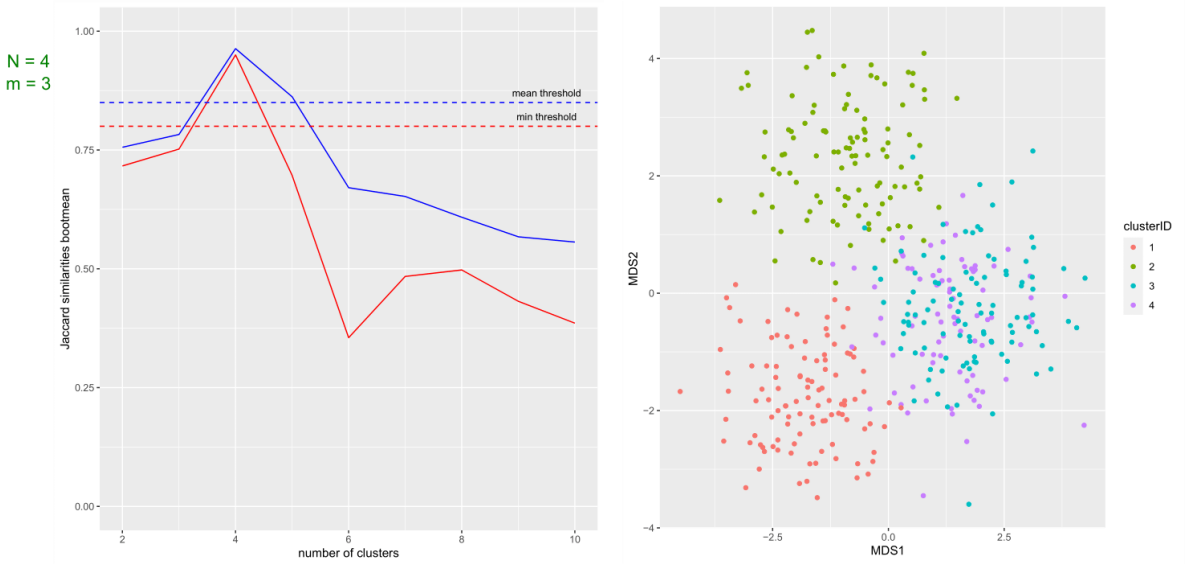


Figure 2.4. Clusterboot double threshold explanation
We show clusterboot output computed with "clusterboot" R function expressed through the mean value of Jaccard similarities matrix (blue line) and the minimum value (red line) in a specified case (N = 4, m = 3) to underline the requirement of the further threshold on minimum: considering only the mean value and the corresponding threshold (blue dashed line, 0.85) the estimated number of clusters is 5 instead to real number 4; also considering the minimum value and the corresponding threshold (red dashed line, 0.80) we obtain the correct estimation of number of clusters.

2.3.3 Accuracies evaluation

Considering the tested scenarios, we found that clusterboot algorithm is able to predict the right number of clusters in 76% of generated simulated data, whereas prediction strength works well in 33%. (**Table 2.2**). Moreover, results showed that, as the number of classes increases, the mean value to compute the distance among classes needs to be increased to determine the clustering.

Table 2.2. Number of clusters estimation in clusterboot compared with prediction strength
We computed prediction strength ("prediction.strength" R function) and clusterboot ("clusterboot" R function) algorithms on simulated data as vary N (number of clusters) and m (value of mean that computes the distance among clusters). We report the number of estimated clusters and the error (no = number of estimated clusters is equal to real number of clusters, yes = otherwise).

		number of estimated clusters		error	
mean among clusters	**#clusters**	**clusterboot**	**prediction strength**	**clusterboot**	**prediction strength**
0	2	2	1	yes	no
0	3	1	1	no	no
0	4	1	1	no	no
0.5	2	2	1	no	yes
0.5	3	1	1	yes	yes
0.5	4	1	1	yes	yes
1	2	2	1	no	yes
1	3	1	1	yes	yes
1	4	1	1	yes	yes
1.5	2	2	1	no	yes
1.5	3	1	1	yes	yes
1.5	4	1	1	yes	yes
2	2	2	1	no	yes
2	3	3	1	yes	yes
2	4	1	1	yes	yes
2.5	2	2	3	no	yes
2.5	3	3	1	no	yes
2.5	4	4	1	no	yes
3	2	2	2	no	no
3	3	3	1	no	yes
3	4	4	1	no	yes
3.5	2	2	2	no	no
3.5	3	3	4	no	yes
3.5	4	4	1	no	yes
4	2	2	2	no	no
4	3	3	4	no	yes

4	4	4	4	no	no
4.5	2	2	2	no	no
4.5	3	3	4	no	yes
4.5	4	4	4	no	no
5	2	2	2	no	no
5	3	3	4	no	yes
5	4	4	4	no	no

2.3.4 Clusterboot works better than prediction strength

The **Fig 2.5** shows that the methodology based on clusterboot works better than prediction strength. Both algorithms work well when classes are generated with a mean value equal to 0, as expected. In this case, the classes are not well separated and the implementation of both algorithms provided a similar result: the curve is below threshold values (**Fig 2.5A**), meaning there are no clusters. In the other cases, we found that methodology based on clusterboot works better than prediction strength. In particular, increasing mean value, clusterboot results predicted the right number of clusters, unlike prediction strength that in some cases is not able to identify the clustering in the data (**Fig 2.5B**) and in other cases prediction is wrong (**Fig 2.5C**).

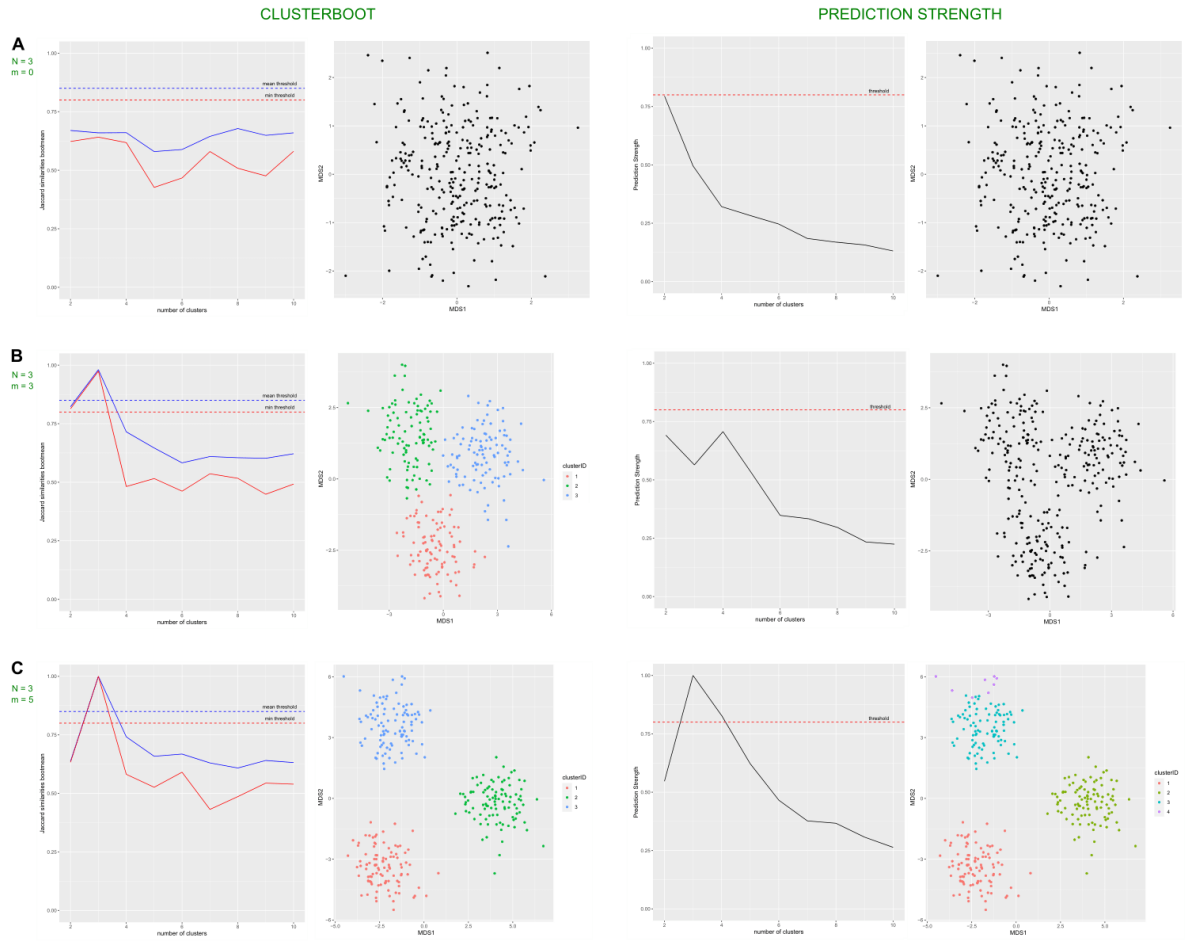


Figure 2.5. Clusterboot compared with Prediction strength
We report clusterboot results computed with "clusterboot" R function (on the left) and prediction strength results computed with "prediction.strength" R function (on the right) and their corresponding MDS obtained with "cmdscale" R function with colored clusters, if any. We show different cases as vary m (value of mean that computes the distance among clusters) and N (number of clusters) is equal to 3: A) For m = 0 there are no clusters and both algorithms work well; B) For m = 3 clusterboot works while prediction strength is unable to identify the clusters; C) For m = 5 clusterboot still works while prediction strength overestimates the number of clusters.

2.3.5 Limit of detection

The methodology based on clusterboot doesn't work in 24% of considered scenarios, particularly when the mean value to compute the distance among classes was close to 0. In these cases prediction strength doesn't work too, except in the generated dataset with 2 classes and a mean value equal to 0, in which clusterboot metric provided as result a clustering. Another limit of our detection depends on the number of classes: as this number increases, it needs a stronger separation among classes, by increasing the mean value, to discriminate among them with our methodology.

2.3.6 Computational time

We evaluated the computational time of both algorithms to show that clusterboot exhibited better performances than prediction strength. Overall, the required time to compute clusterboot for the assessment of the number of clusters was less than prediction strength computational time and therefore the methodology based on clusterboot can potentially be applied in large-scale scenarios. For the assessment of computational time, we performed clusterboot and prediction strength on different data varying the number of samples/observations in the range [10:1000] with a step value equal to 5 for clusterboot and 45 for prediction strength to determine the computational time in each configuration. The results showed that the required time to compute both increases as the number of samples/observations increases, but clusterboot is high-performing in terms of computational time (**Figure 2.6A**). In particular, clusterboot computational time was less than 1 second when applied on small-size data (∼100 samples) and about 25 seconds for a dataframe with 1,000 samples/observations, that is less than 1 minute. Prediction strength required higher computational time than clusterboot (red line in **Figure 2.6B**). Specifically, it needed less than 1 minute when applied on small-size data (∼100 samples) and over 2 hours for a dataframe with 1,000 samples/observations. The difference between the tested algorithms in terms of computational time was greater as the number of observations increases, as shown in **Figure 2.6B**.

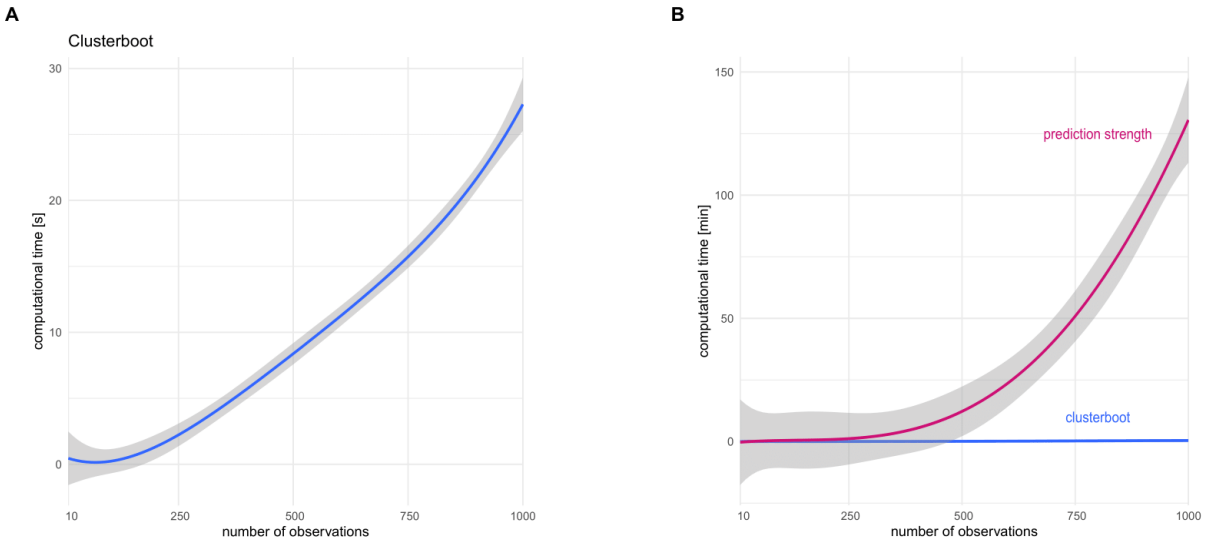


Figure 2.6. Computational time estimation
The computational time of both estimated algorithms increases as the number of observations increases. We assessed the required time as vary the number of observations in the range [10:1000]. A) The clusterboot (computed with "clusterboot" R function) were applied varying the number of observations from 10 to 1,000 with a step value equal to 5. The required time is expressed in seconds, whereas for (B) the prediction strength (computed with "prediction.strength" R function) is expressed in minutes. In this case the number of observations was varied from 10 to 1,000 with a step value equal to 45. We also reported clusterboot results (blue line) in the same plot, in order to properly show the better performances of clusterboot in terms of computational time.

2.4 Discussion and Conclusions

In this chapter, we validated a methodology aiming at estimating in an automatic way the number of clusters given a distance matrix. We compared two methodologies based on prediction strength and clusterboot. While we considered hierarchical clustering and clara algorithms for prediction strength and clusterboot, respectively, other clustering methods could be considered to further improve the methodology. Actually, preliminary analysis using PAM as clustering method for prediction strength gave unsatisfactory results in terms of prediction of the true number of clusters and were therefore not further investigated.

The two methodologies were mainly validated on synthetic data. Such data were generated by varying the number of clusters and correspondingly the number of features. For simplicity, the number of features was equal to the number of classes, and by considering Gaussian random distribution with standard deviation fixed to 1. Also the number of observations was kept constant for all clusters. This preliminary analysis could be improved by generating and testing

more scenarios by changing these settings. Similarly, also other metrics for estimating distances could be considered; while we chose the Euclidean distance, other distances could be evaluated.

We finally observe that the methodology based on clusterboot required a further threshold on the Jaccard similarity values than what reported in literature [42], empirically defined. While this may make the estimation process more complex, it actually showed improved performances in the different testing settings.

Extension from synthetic to real data validation was performed for the identification of sub-species in microbial species of great relevance in the food science and human microbiome fields. We will show in Chapter 4 a large-scale analysis performed on the well known *Bifidobacteriaceae* family. In particular, we estimated in different prevalent *Bifidobacteriaceae* species a number of sub-clades equal to the number of known subspecies in the literature and these sub-clades were coincident with them. This result underlines that, although we validated our methodology on synthetic data, it also works well in real scenarios.

3 Lactic acid bacteria diversity in food and human microbiomes from large-scale metagenomic analysis

3.1 Introduction and scientific rationale

For several decades, lactic acid bacteria (LAB) have been among the most extensively studied microorganisms. LAB have a fundamental role in different biological processes and ecosystems, especially with respect to fermented foods. The microbiology of fermentations has been extensively studied for over a century and the ability to transform raw materials into edible products with defined characteristics dates back to thousands of years as a strategy of food preservation [44,45]. Industrial fermentations are based on selected cultures that are used as starters or adjuncts to guarantee specific metabolic activities along with quality, reproducibility, and safety. On the other hand, artisanal processes do not usually involve defined starter cultures and the LAB available in the raw materials, or sourced from a previous manufacture, lead the fermentation. Food-associated LAB have been studied mainly from the perspective of their fermentation performances and phenotypic properties, and knowledge on such properties has recently increased thanks to intense genome sequencing of LAB strains [46,47].

Apart from their contributions to food quality and safety, LAB have attracted considerable interest due to their potentialities to add functional properties to certain foods or as supplements. Functional foods are designed to deliver additional benefits over their basic nutritional values and contribute to human health [48]. In this regard, several LAB species and strains have been recognized as probiotics, i.e., “live microorganisms that confer a health benefit on the host when administered in adequate amounts” [49]. Importantly, many LAB species also enjoy generally recognized as safe (GRAS) status.

Despite the extensive literature focusing on characterizing LAB in food, it is still not fully understood how they interact with the human gut microbiome [48]. Ingested LAB need to first survive the physical and chemical barriers of the gut, before competing with hundreds of different species, and finally being able to exert their beneficial effects. Indeed, LAB are regarded as components of the transient gut microbial community, coming from the external

environment and with food representing the main source, which interacts daily with the longer term members of the gut microbiome [50]. Despite this general view, it is still not known to what extent components of the food microbiome are actively transferred to become part of the gut microbiome and what role they play in this complex environment. Depending on the specific food, technology of production and fermentation process, fermented foods can harbour several LAB species and strains and are natural sources of live microorganisms that are consumed daily across all human populations and that can potentially interact with the gut microbiome. Despite this, the degree to which LAB species and strains not explicitly regarded as probiotics can be transferred to the gut has been largely underexplored. Additionally, no studies have been conducted to assess the distribution of LAB in the global population, a gap that may be bridged by taking advantage of the growing availability of high throughput sequencing data.

In this chapter, we perform a large-scale genome-wide analysis of publicly available and newly sequenced food and human metagenomes to investigate the prevalence and diversity of LAB species with a view to identifying links between gut and food microbiomes. We find that LAB species occur with variable prevalence and generally low abundance in the human gut. Such prevalence is affected by age and lifestyle. LAB species identified in food only partially match those in the gut. Comparative genomics suggest an overall food origin for the gut strains.

3.2 Materials and methods

3.2.1 Publicly available and newly acquired food metagenomes

We considered and curated public datasets from fermented food metagenomes in addition to food metagenomes newly sequenced in this study. In total we put together 303 samples spanning 11 datasets and coming from different types of cheese (N = 191), fermented foods (N = 58), nunu (N = 20), milk kefir (N = 18), and yogurt and dietary supplements (N = 16) [51–57]. More information is detailed in **Table 3.1**. Additional information on the newly acquired metagenomes is available in **Supplementary Table 3.1**.

3.2.2 Publicly available human metagenomes

In addition, we considered publicly available metagenomic datasets corresponding to the human microbiome. More specifically, we included 47 human microbiome datasets totalling

9,445 metagenomes and 4.2e11 Illumina reads as done in [24] (seventeen metagenomes that were left out due to technical issues in [24] were included here by marginally expanding the original set of 9,428 metagenomes). Overall, the samples were acquired from six major body sites: the gut by stool sampling (N = 7,907), oral cavity (N = 785), skin (N = 508, including from the anterior nares), airways (N = 151), vagina (N = 86), and breast milk (N = 8, data not included in figures). These samples covered 31 countries that were grouped by continent as follows: Africa (MDG: Madagascar, TZA: Tanzania), Asia (BGD: Bangladesh, BRN: Brunei, IDN: Indonesia, ISR: Israel, KAZ: Kazakhstan, MNG: Mongolia, MYS: Malaysia, SGP: Singapore), China (CHN, which we kept separated from the other Asian countries due to its large sample size), Europe (AUT: Austria, DEU: Germany, DNK: Denmark, ESP: Spain, EST: Estonia, FIN: Finland, FRA: France, GBR: Great Britain, HUN: Hungary, ISL: Iceland, ITA: Italy, NLD: The Netherlands, NOR: Norway, RUS: Russia, SVK: Slovakia, SWE: Sweden), North America (CAN: Canada, USA: United States), Oceania (FJI: Fiji), and South America (PER: Peru). The samples were also categorized as corresponding to westernized (N = 8,850) and non-westernized (N = 595) lifestyles [24]. More specifically, westernization is a complex process that occurred during the last few centuries and that involved lifestyle changes compared to populations prior to the modern era. Such changes include increased hygiene and sanitized environments, introduction of antibiotics and other drugs, increased high-calorie high-fat dietary regimes, enhanced exposure to pollutants, and reduced contact with wildlife and domesticated animals. We adopt westernized and non-westernized as umbrella terms to depict populations that differ by the majority of the aforementioned factors even though this definition comprises heterogeneous populations. Finally, these metagenomes spanned multiple age categories: newborns (N = 711, < 1 year of age), children (N = 802, age ≥1 and <12 years), school age individuals (N = 215, age ≥12 and <19 years), and adults (N = 7,669, age ≥19). Despite curation efforts, age category metadata corresponding to 48 samples could not be sourced. These manually-curated metadata are available in the **Supplementary Table 3.2** and in the *curatedMetagenomicData* package [58].

3.2.3 Taxonomic profiling of food and human metagenomes

Quantitative taxonomic profiling was applied on the 9,445 human metagenomes and the 303 food metagenomes by applying MetaPhlAn3 [59] with default parameters. MetaPhlAn3 estimates relative abundances of microbial species using the pre-generated ~1M unique clade-specific marker genes identified from ~17,000 reference genomes (~13,500 bacterial and

archaeal, ~3,500 viral, and ~110 eukaryotic). Taxonomic profiles along with associated metadata information are available in **Supplementary Table 3.2**. We detected 152 species belonging to the Lactobacillales order occurring in at least one of the metagenomes with a relative abundance greater than 0.01%. Among them, we identified 70 species belonging to the LAB group (i.e., species belonging to *Lactobacillus*, *Lactococcus*, *Leuconostoc*, and *Weissella* genera in addition to *S. thermophilus*), and restricted the rest of the analysis to the 30 of them having a prevalence greater than 0.1% in the human gut. Taxonomic profiles of these 30 species are reported in **Figure 3.1** and **Supplementary Figures 3.1-3.4**. Prevalence was computed by thresholding relative abundance at 0.01%. Average relative abundance was computed on positive samples only.

3.2.4 Metagenome-assembled genomes (MAGs) reconstruction

Taxonomic profiling was coupled with the reconstruction of microbial genomes directly from metagenomes. The approach that we validated in [24] was applied here to reconstruct metagenome-assembled genomes (MAGs) MAGs from food metagenomes. More specifically, single-sample metagenomics assemblies were generated with metaSPAdes [60] (version 3.10.1; default parameters) or IDBA-UD [61] (version 1.1.3; default parameters). Contigs longer than 1,000 nt were binned with MetaBAT2 [62] (version 2.12.1; option '-m 1500'). Quality control with CheckM (v. 1.0.7) [63] yielded 666 medium-quality food MAGs (completeness > 50% and contamination <5%) of sufficient quality according to previous recommendations [64]. These newly reconstructed MAGs were then considered within the human MAG catalogue totaling 154,723 MAGs reconstructed from the 47 human datasets considered in this study [24].

3.2.5 Clustering of genomes into species-level genome bins (SGBs)

The 155,389 MAGs described in the previous section were integrated with the set of 193,078 reference genomes available in GenBank as of March 2019. This resulted in a total of 348,467 genomes that were clustered into species-level genome bins (SGBs) following the procedure proposed in [24]. Genomes were clustered with average linkage at 5% genetic distance based on whole-genome nucleotide similarity estimation using Mash (v. 2.0; option "-s 10000" for sketching) [33]. The 666 food MAGs were grouped by this procedure into 171 SGBs: 108

SGBs (comprising 574 MAGs) contained at least one reference genome or human MAG (kSGBs), while a further 63 SGBs (comprising 92 MAGs) consisted only of genomes reconstructed in this study from food metagenomes (fSGBs). Summaries of the newly generated MAGs and SGBs are available in **Figure 3.2A**, **Figure 3.2B**, **Supplementary Table 3.5** and **Supplementary Table 3.6**.

3.2.6 Metadata curation for selected LAB species

We considered the 30 selected LAB species shown in **Figure 3.1** for comparative genomics purposes. Among the 348,467 genomes described in the previous section, 2,859 genomes (comprising 1,042 MAGs) were included in SGBs containing at least one reference genome assigned to these 30 species and were kept for further analyses. We retrieved and manually curated the source type in all cases. For reference genomes, the source of isolation was extracted from the NCBI portal or from related publications. Genomes were grouped in three categories based on the source type: "human", "food", and "other". Genomes for which this information was missing were labelled as "NA" (N = 226, 7.9% of the cases). More information relating to these 2,859 genomes is available in **Supplementary Table 3.7**.

3.2.7 Reconstruction of phylogenetic structure

Phylogenies were built using the newly developed PhyloPhlAn 3.0 package that extends the original PhyloPhlAn2 version [65]. Each SGB-specific phylogeny (**Figure 3.3**) was based on the set of species-specific marker genes that can be retrieved in PhyloPhlan 3.0 with the command phylophlan2_setup_database.py. The number of marker genes for each SGB is summarized in **Supplementary Table 3.10**. This departs from the default option in using the 400 universal markers available in PhyloPhlAn 3.0 and guarantees a higher resolution of the built phylogenies. The parameters were set as follows "--diversity low --fast --min_num_marker 50", which indicated that genomes mapping less than 50 markers were discarded from the phylogeny. External tools embedded in PhyloPhlan 3.0 were run with their specific options as follows:

- blastn (version 2.6.0+; [66]) with parameters "-outfmt 6 -max_target_seqs 1000000"
- mafft (version 7.310; [67]) using the "L-INS-i" algorithm and with parameters "--anysymbol --auto"

- trimal (version 1.2rev59; [68]) with parameter "-gappyout"
- FastTree (version 2.1.9; [69]) with parameters "-mlacc 2 -slownni -spr 4 -fastest -mlnni 4 -no2nd -gtr -nt"
- RAxML (version 8.1.15; [70]) with parameters "-p 1989 -m GTRCAT -t <phylogenetic tree computed by FastTree>"

Phylogenetic trees (**Figure 3.3** and **Figure 3.4**) were visualized with GraPhlAn [71]. Additionally, multidimensional scaling (MDS) plots (**Figure 3.3**, **Figure 3.4**, **Supplementary Figure 3.4**, and **Supplementary Figure 3.5**) were built on the whole-genome Average Nucleotide Identity (ANI) distances computed with FastANI [32].

3.2.8 Functional analysis and statistical significance

The set of genomes (MAGs and reference genomes) considered in this study was annotated with Prokka (v. 1.12; [72]) using default parameters. Proteins inferred by Prokka were then processed with Roary [73] (v. 3.11; option '-i 90') to generate the presence-absence binary matrix on the core and accessory genes. Gene enrichment within human and food genomes was determined by considering only MAGs and reference genomes having completeness >80% in order to avoid possible biases coming from highly incomplete genomes and by taking into account genes present in at least 5% and less than 95% of the genomes. Statistical significance was tested through Fisher's test with false discovery rate (FDR) correction for multiple hypothesis testing.

3.3 Results

3.3.1 Large-scale meta-analysis on food and human microbiomes

We performed a large-scale meta-analysis on microbiomes from food sources and human body sites to investigate the prevalence and diversity of LAB species in the human microbiome and their overlap with species and strains found in food. To achieve this goal, we considered 303 food metagenomes (152 publicly available and 151 obtained in this study) (11 datasets, **Table 3.1** and **Supplementary Table 3.1**) that we curated in this study, which corresponded to different types of fermented foods and beverages [51–57]. In addition, we considered 9,445

human metagenomes from 47 public datasets spanning multiple body sites (84% from the gut), age categories, countries, and lifestyles that we retrieved from recent meta-analyses [24,58].

Table 3.1. Summary of the analysed food metagenomic datasets.

Study	Type of food	# samples	Accession number	Reference
BertuzziAS_2018	Surface ripened cheese	42	PRJEB15423	[51]
Escobar-ZepedaA_2016	Mexican ripened cheese	1	PRJNA286900	[52]
LeechJ_2019	Fermented food	58	PRJEB35321	This study
MacoriG_2019	Cheese	77	PRJEB32768	This study
MilaniC_2019	Parmesan cheese	2	PRJNA482503	[53]
PasolliE_2019	Yogurt and dietary supplement	16	PRJNA603575	This study
PfeferT_2018	Cheese	36	PRJNA430402	-
QuigleyL_2016	Continental type cheese	10	PRJEB6952	[54]
WalshAM_2016	Milk kefir	18	PRJEB15432	[55]
WalshAM_2017	Nunu	20	PRJEB20873	[56]
WolfeBE_2014	Smear ripened cheese	23	mgp3362	[57]

3.3.2 Variable prevalence of LAB in the human gut

We considered reference-based taxonomic profiles [59] of all 9,445 human metagenomes [24,58] (see **Methods**) and focused specifically on LAB species in this study (**Supplementary Table 3.2**). We detected 152 species belonging to the Lactobacillales order occurring in at least one of the metagenomes with a relative abundance greater than 0.01%. Among them, we identified 70 species belonging to the LAB group, and restricted the following analysis to the 30 of them having a prevalence greater than 0.1% in the human gut (see **Methods**). These represented mainly species (spanning *Lactobacillus*, *Lactococcus*, *Leuconostoc*, *Streptococcus*, and *Weissella* genera) of potential food origin, including bacteria occurring in probiotic supplements, in addition to typically non-food origin species such as *Lb. mucosae*, *Lb. ruminis*, and *Lb. salivarius* (**Figure 3.1**). The two most prevalent species in the gut were *Streptococcus thermophilus* (prevalence 31.2%, i.e., present at > 0.01% relative abundance in 31.2% of the gut metagenomes) and *Lc. lactis* (16.3%), both commonly found in dairy products (**Figure 3.1**,

Supplementary Figure 3.1, and **Supplementary Table 3.3**). Multiple *Lactobacillus* species of predominantly food origin were detected at lower prevalence (3%-5%) and comprised *Lb. casei/paracasei*, *Lb. delbrueckii*, *Lb. fermentum*, and *Lb. rhamnosus*). Non-food origin bacteria were also identified at remarkable levels such as *Lb. ruminis* (11.0%), *Lb. salivarius* (4.7%), and *Lb. mucosae* (4.0%). While prevalence was variable, average relative abundance (computed on positive samples only) of single species was generally rather low (<2%), including the case of the two most prevalent species *S. thermophilus* (0.6%) and *Lc. lactis* (0.4%). Exceptions (rel. ab. >2%) were verified for *Lb. amylovorus*, *Lb. brevis*, and *Lb. buchneri*, which however rarely occurred (prev. < 1%).

Strong age-related patterns were verified for some of the species prevalent in gut samples (N = 7,907) (**Figure 3.1**, **Supplementary Figure 3.2**, and **Supplementary Table 3.4**). *S. thermophilus* increased in prevalence from newborns (8.4%) to adults (33.7%, $p < 1e\text{-}40$), with comparable average abundance. This may reflect the increase in consumption of yogurts and other dairy products that can be sources of *S. thermophilus [74]*. A similar pattern was observed for *Lb. delbrueckii* ($p < 1e\text{-}10$) and the non-food origin species *Lb. mucosae* ($p < 1e\text{-}10$), *Lb. ruminis* ($p < 1e\text{-}20$), and *Lb. salivarius* ($p < 1e\text{-}10$), which suggests their gut colonization later in age. Also, *Lc. lactis* had higher prevalence in adults (15.8%) than newborns (8.6%, $p < 1e\text{-}6$), with its detection in only one infant cohort originating from Estonia, Finland, and Russia [75]. Other lactobacilli were more prevalent and abundant in newborns such as *Lb. casei/paracasei* ($p < 1e\text{-}20$ with respect to adults), *Lb. gasseri* ($p < 1e\text{-}7$), *Lb. plantarum* ($p < 1e\text{-}4$), and *Lb. rhamnosus* ($p < 1e\text{-}70$). These species have also been detected in human breast milk [76] suggesting their possible transmission from mother to infant through breastfeeding, as previously reported for *Lb. plantarum* [77]. Notably, these species were not found to be vertically transmitted from other mother's body sites[78].

Overall, we found that LAB are a subdominant component of the gut microbiome, although several species exhibited non-negligible contributions. More specifically, we identified twenty-one LAB occurring with prevalence greater than 1% and eighteen with relative abundance greater than 0.5% when detected in the gut. It is reasonable to hypothesize that those species may be short- or long-term colonizers of the human microbiome.

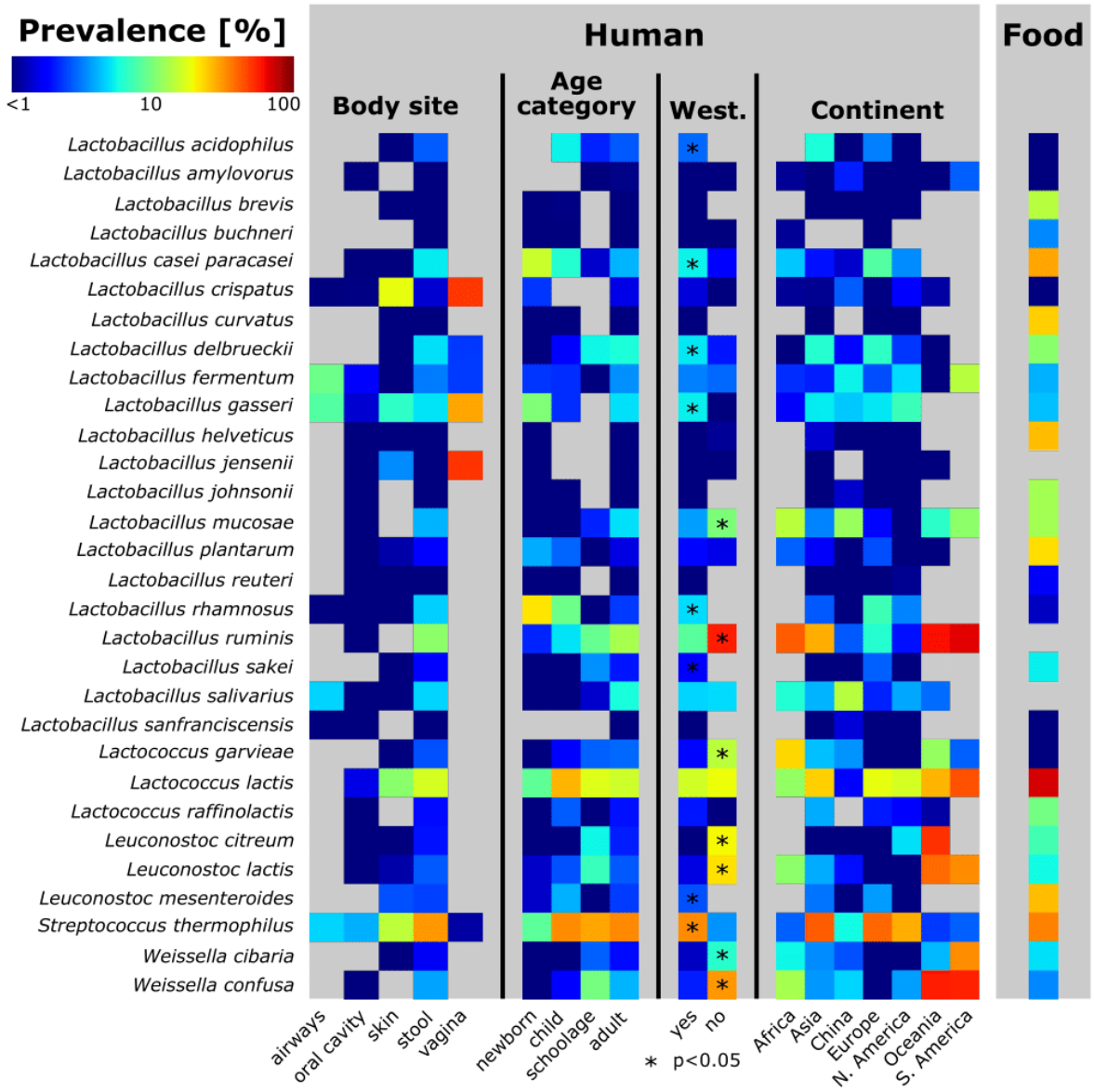


Figure 3.1. Average prevalence of LAB species from human and food microbiomes.
We report the 30 LAB species having a prevalence greater than 0.1% in the human gut. Values are obtained from 9,445 publicly available human metagenomes and stratified by multiple host conditions (i.e., body site, age category, westernized lifestyle, and continent). Age category, westernized lifestyle, and continent statistics refer to stool samples only. Food results are obtained from 303 food metagenomes. Numbers and p-values (Fisher's test, false discovery rate correction) in Supplementary Figures 3.1-3.4 and Supplementary Table 3.4. Relative abundances in Supplementary Table 3.2 and Supplementary Table 3.3.

3.3.3 Occurrence and abundance of LAB is linked to lifestyle

We then stratified the gut metagenomes in terms of host lifestyles (**Figure 3.1**, **Supplementary Figure 3.3**, and **Supplementary Table 3.4**), which revealed variations in prevalence and

abundance between westernized and non-westernized populations for multiple species. Higher prevalence in westernized populations was observed for six lactobacilli, mostly of food origin, such as *Lb. acidophilus* ($p < 1e-6$), *Lb. casei/paracasei* ($p < 1e-4$), *Lb. delbrueckii* ($p < 0.01$), *Lb. gasseri* ($p < 1e-6$), *Lb. rhamnosus* ($p < 1e-9$), and *Lb. sakei* ($p < 1e-3$). By contrast, *Lb. mucosae* ($p < 1e-8$) and *Lb. ruminis* ($p < 1e-100$) that do not occur in food were more prevalent in the non-westernized cohorts. Despite different patterns in terms of prevalence, all lactobacilli were on average more abundant in the westernized populations. Among the other genera, *S. thermophilus* was highly prevalent in the westernized cohorts ($p < 1e-50$). Higher prevalence in the non-westernized group was observed for *Lactococcus garvieae* ($p < 1-e30$) in addition to multiple heterofermentative species such as *Leuconostoc citreum* ($p < 1e-70$), *Leuconostoc lactis* ($p < 1e-60$), *Weissella cibaria* ($p < 1e-10$), and *Weissella confusa* ($p < 1e-100$), which is consistent with their widespread prevalence in raw vegetables [79] that are likely consumed in such populations. In fact, non-western populations usually have hunter-gatherer diet and lifestyle, which is recognized to be characterized by high consumption of tubers, drupes, roots, and fruits [80,81]. Indeed, it was also reported that the !Kung and the Hadza, two non-Western African populations, still obtain 60–80% and 50–65% of their diet from plant foods, respectively [82].

We further grouped metagenomes by host country of origin (see **Methods**) and identified more subtle geographical variations (**Figure 3.1** and **Supplementary Figure 3.4**). Overall, food-associated lactobacilli were most prevalent and abundant in Europe, were less so in Asia and North America, and almost absent in China (kept distinct from the other Asian countries due to its large sample size) and in the non-westernized populations. The higher prevalence in European cohorts was significant ($p < 0.05$) for *Lb. casei/paracasei* (8.0%), *Lb. delbrueckii* (6.6%, with a similar value in Asia), and *Lb. rhamnosus* (7.1%). Exceptions were *Lb. gasseri*, having comparable prevalence in continents including westernized cohorts, and *Lb. fermentum*, more prevalent in North America, South America, and China, with the latter observation being consistent with its widespread occurrence in Chinese fermented foods [83]. Non-food lactobacilli were not prevalent in Europe. *Lb. mucosae* exhibited high prevalence (>10%) in Africa, China, and South America, with comparable abundance across the globe. A similar trend was verified for *Lb. ruminis*, although with higher prevalence in non-westernized cohorts, while the presence of *Lb. salivarius* was distinctive for the Chinese population ($p < 0.01$). Among the other genera, *Lc. lactis* exhibited high prevalence across the entire globe (ranging from 11.5% in Africa to 44.4% in South America) with the sole exception of China (1.7%). *S.*

thermophilus reached high prevalence in Asia (41.5%), Europe (39.6%), and North America (28.1%), but was much less prevalent in the Chinese (5.6%) and non-westernized (<3%) cohorts.

3.3.4 LAB species from food only partially match those in the gut

We established genome level links between the microorganisms populating the human microbiome and those found in food by integrating the genomes reconstructed from a set of 9,445 human metagenomes with those from the set of 303 food metagenomes that we generated, collected and curated in this work (**Table 3.1** and **Supplementary Table 3.1**). More specifically, we considered 303 metagenomic samples spanning eleven datasets and coming from different types of cheese (N = 191), multiple fermented foods (N = 58), nunu (N = 20), milk kefir (N = 18), and yogurt and dietary supplements (N = 16). We applied a validated [24,84] computational pipeline that combined single-metagenome assembly, contig binning, and genome quality control to reconstruct *de novo* metagenome-assembled genomes (MAGs) from the set of food metagenomes (see **Methods**). We generated a total of 666 food MAGs (completeness > 50% and contamination <5%) of sufficient quality according to previous recommendations [64]. These MAGs from food were integrated with the set of 154,723 MAGs that we retrieved from the 9,445 human metagenomes using the same assembly-based pipeline [24] and with the set of 193,078 reference genomes (available in GenBank as of March 2019). This resulted in a total of 348,467 genomes that were clustered at 5% genetic distance based on whole-genome nucleotide similarity estimation and recapitulated in species-level genome bins (SGBs, i.e., clusters of genomes spanning 5% genetic diversity, see **Methods**). The 666 food MAGs were grouped into 171 SGBs (**Supplementary Table 3.5** and **Supplementary Table 3.6**), which we discuss below on the basis of their occurrence in food samples and human gut (**Figure 3.2A** and **Figure 3.2B**).

Most of the food MAGs (349, 52.4%) belonged to SGBs also found in the human gut, with 265 of them associated with twenty of the thirty LAB species discussed previously (**Figure 3.2A, top panel** and **Supplementary Figure 3.5**). The species most reconstructed from food sources was *Lc. lactis* (N = 90 MAGs), with 86 MAGs extracted from cheese. Sixty MAGs were associated with *S. thermophilus*, the majority of them was reconstructed from cheese and yogurt, and five additional genomes were extracted from different fermented foods such as wagashi, beetroot kvass, ryazhenka, ruž'a, and labne. A consistent number of MAGs was also

retrieved from *Lb. helveticus* (33 MAGs from cheese), *Lb. curvatus* (14 MAGs from cheese and one from sauerkraut), *Lb. delbrueckii* (11 MAGs from cheese or yogurt in addition to single genomes from dietary supplement and tofu), *Leuconostoc mesenteroides* (5 MAGs from nunu and single genomes from bread kvass, ginger beer, milk kefir, beetroot kvass, ruž'a, and cheese), and *Lb. casei/paracasei* (4 MAGs from cheese, 2 MAGs from dietary supplements, and 2 MAGs from water kefir). We also extracted 4 MAGs of *Lb. mucosae*, a typically non-food microorganism that is usually found in the intestine of pigs or other animals [85] and that we instead reconstructed from different fermented foods such as kimchi, kombucha vinegar, agousha, and sauerkraut.

We identified seventeen additional non-LAB SGBs having MAGs from both food and human metagenomes, for a total of 84 food MAGs (12.6%; **Figure 3.2A, bottom panel**) and spanning three phyla (namely Actinobacteria, Firmicutes, and Proteobacteria). Some of these may be microbial contaminants in the food chain that can arise from different sources including animal, feed, and soil [86,87]. The SGB with the most MAGs (N = 16) was that containing *Streptococcus equinus* and *Streptococcus infantarius* genomes, two species usually found in the rumen [88] but occasional pathogens for humans [89] and that we found in African fermented foods [56].

The majority of the food SGBs (134 out of 171), accounting for 317 MAGs (47.6%), did not exhibit an overlap with human MAGs, likely representing species unable to reach the colon or characterized by low prevalence and abundance in the human gut (**Figure 3.2B**). Among them, 71 SGBs (53.0%; comprising 225 MAGs) contained at least one reference genome (kSGBs; **Figure 3.2B, left panel**). The most prevalent food-specific species was *Brevibacterium linens* (24 MAGs) which was reconstructed from multiple cheese types (i.e., surface ripened [51], smear ripened [57], hard, and tomme). Food-specific SGBs also included *Staphylococcus saprophyticus* (13 MAGs), *Glutamicibacter arilaitensis* (12 MAGs), and 58 MAGs from 21 LAB species spanning 6 families, the most prevalent being *Lc. lactis* subsp. *cremoris*. This set of MAGs and reference genomes showed a >5% genetic distance from *Lc. lactis* subsp. *lactis* genomes [90], which we kept as a separate SGB (ID 7985) and found to be prevalent in both food and human metagenomes, in contrast to *Lc. lactis* subsp. *cremoris* which was only detected in food metagenomes. Similarly, *Lc. raffinolactis* was divided into two SGBs, with human and food MAGs grouped in the SGBs 7989 and 7991, respectively.

Out of the 134 SGBs not overlapping with human MAGs, 63 SGBs (47%; comprising 92 MAGs) consisted of MAGs reconstructed in this study from food metagenomes without any

reference genomes. These represented new species currently not represented in public repositories (**Figure 3.2B, right panel**), of which only 12 were assigned to known genera, and which should be targeted for cultivation-based analysis.

The set of genomes reconstructed and the SGBs identified in this study and that we made publicly available (see **Methods**) facilitated a more in-depth comparative genomics analysis.

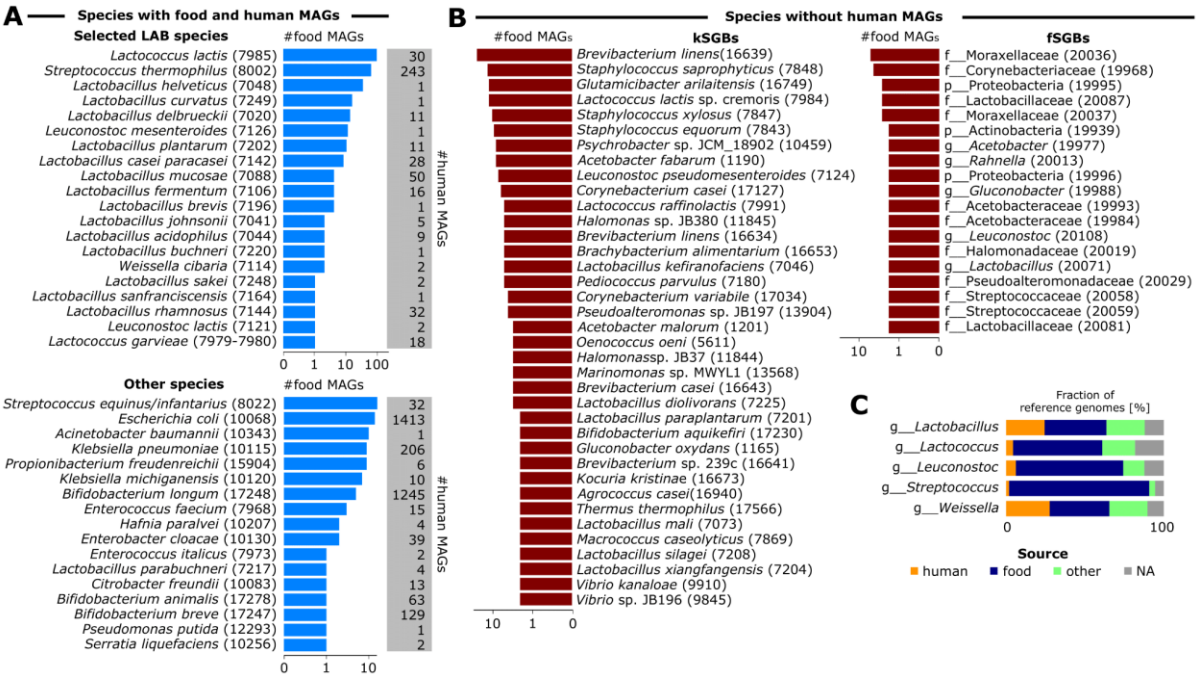


Figure 3.2. Microbial genomes reconstructed from food metagenomes.
A) Most prevalent species-level genome bins (SGBs) in 666 MAGs reconstructed from 303 food metagenomes and overlapping with human MAGs (i.e., found in at least one of the 154,723 human MAGs). Numbers in parenthesis represent the SGB IDs. **B)** Most prevalent food SGBs not overlapping with human MAGs. kSGBs denote SGBs with at least one reference microbial genome, whereas fSGBs identify newly assembled SGBs from food metagenomes only. X-axes for panels **A)** and **B)** are in logarithmic scale. **C)** Fraction of reference genomes per source type for the 30 selected LAB species and grouped by genera (the same plot at species-level is reported in **Supplementary Figure 3.6**). Raw data in **Supplementary Table 3.6** and **Supplementary Table 3.7**.

3.3.5 Comparative genomics suggests a food origin for the gut strains

Within the available set of MAGs and reference genomes, we performed strain-level comparative genomic analysis for the set of 348,467 genomes previously described and comprising 193,078 reference genomes, 154,723 human MAGs, and 666 food MAGs. The 2,859 genomes (including 1,042 MAGs) associated with the thirty LAB species of interest were

kept for comparative genomics purposes. To inform the comparative analysis, we retrieved and manually curated the source types for all genomes (see **Methods**), and grouped MAGs and reference genomes in three categories: human, food, and other. Genomes for which this information was missing were labelled as NA (7.9% of genomes; **Figure 3.2C**, **Supplementary Figure 3.6** and **Supplementary Table 3.7**).

Overall, two thirds of the reference genomes came from food (43.8%) and human sources (21.0%). The group of genomes from strains not isolated from foods or humans (22.8%) comprised 67 genomes from probiotics and dietary supplements in addition to 347 genomes mainly coming from animal sources. The proportions of species assigned to the different source types was quite variable across species, with a general under-representation of human genomes corresponding to LAB that were prevalent in non-westernized cohorts (**Figure 3.2C**, **Supplementary Figure 3.6**). This reflected the overall scarce availability of genome from isolates for a substantial fraction of the non-pathogenic, commensal members of the human microbiome as recently highlighted [24,91,92]. Reference genomes from human samples were surprisingly almost absent in the case of prevalent species such as *Lc. lactis* (with only one reference genome from the vagina and one MAG from the gut) and *S. thermophilus* (with only one MAG from the gut). The absence of good reference genomes in public repositories prevented the comparison of food and human strains until now, which we aimed to overcome in the present study through an extensive comparative genomics analysis.

S. thermophilus was the species of LAB most frequently reconstructed from metagenomes (243 human and 60 food MAGs; **Figure 3.3A**), an observation consistent with its high prevalence from mapping-based taxonomic profiling (**Figure 3.1**). Comparative genomics, also including 44 reference genomes, did not highlight food-specific or gut-specific sub-clades suggesting that food can be regarded as the main source of this species in the human microbiome. *S. thermophilus* also appeared to be a quite genetically diverse species both in food and human sources with MAGs reconstructed from Asian gut metagenomes enriched in a specific clade (Clade A, **Figure 3.3A**, $p < 1e\text{-}10$). *Lb. delbrueckii* was not prevalent in the gut, and the only two subspecies found in human samples were subsp. *lactis* and subsp. *bulgaricus* (**Figure 3.4A**). Human MAGs of both subspecies clustered together with food MAGs and isolates, again indicating food as the most likely source of this species in the gut. On the other hand, subsp. *delbrueckii*, subsp. *sunkii*, and subsp. *jakobsenii* were found in food, but never reconstructed from the gut. Although *Lb. rhamnosus* was the LAB species for which the greatest number of genomes corresponding to human isolates ($N = 105$) was available, we

collected only 32 human MAGs, which is in agreement with its low prevalence and abundance in the gut (**Figure 3.4B**). We identified a specific cluster including 17% of the *Lb. rhamnosus* human genomes that included the reference genome associated with the *Lb. rhamnosus* strain GG (LGG), which may be due to recent consumption of commercial products due to its wide use in probiotic supplements [93].

The highest number of food MAGs was obtained for *Lc. lactis* (N = 90, **Figure 3.3B**). We refer here to subsp. *lactis*, while subsp. *cremoris* was associated with 12 food MAGs but never reconstructed from human metagenomes. *Lc. lactis* subsp. *lactis* formed two distinct clusters including both food and human genomes. The first cluster included 63% of the genomes, exhibited an overall low diversity (<0.8% genetic distance between closest genome pairs), and included all the food genomes related to cheese and dairy fermentation. The second cluster was more diverse, dominated by environmental and raw vegetable products, and included the only MAG from human skin and the three gut MAGs from non-westernized cohorts. An additional cluster containing two genomes from nunu[56] was never found in humans and exhibited a >3% genetic diversity from all other genomes. Such results highlighted the overall importance of conducting strain-level analysis on the food-gut axis, depicted here by the identification of two main clusters in the human gut associated with different food sources (i.e., one from cheese and dairy fermentation, and the other one from environmental and raw vegetables products). Strains of these clusters are likely characterized by differences in functional traits and potential interaction with the host that deserve to be investigated in future studies.

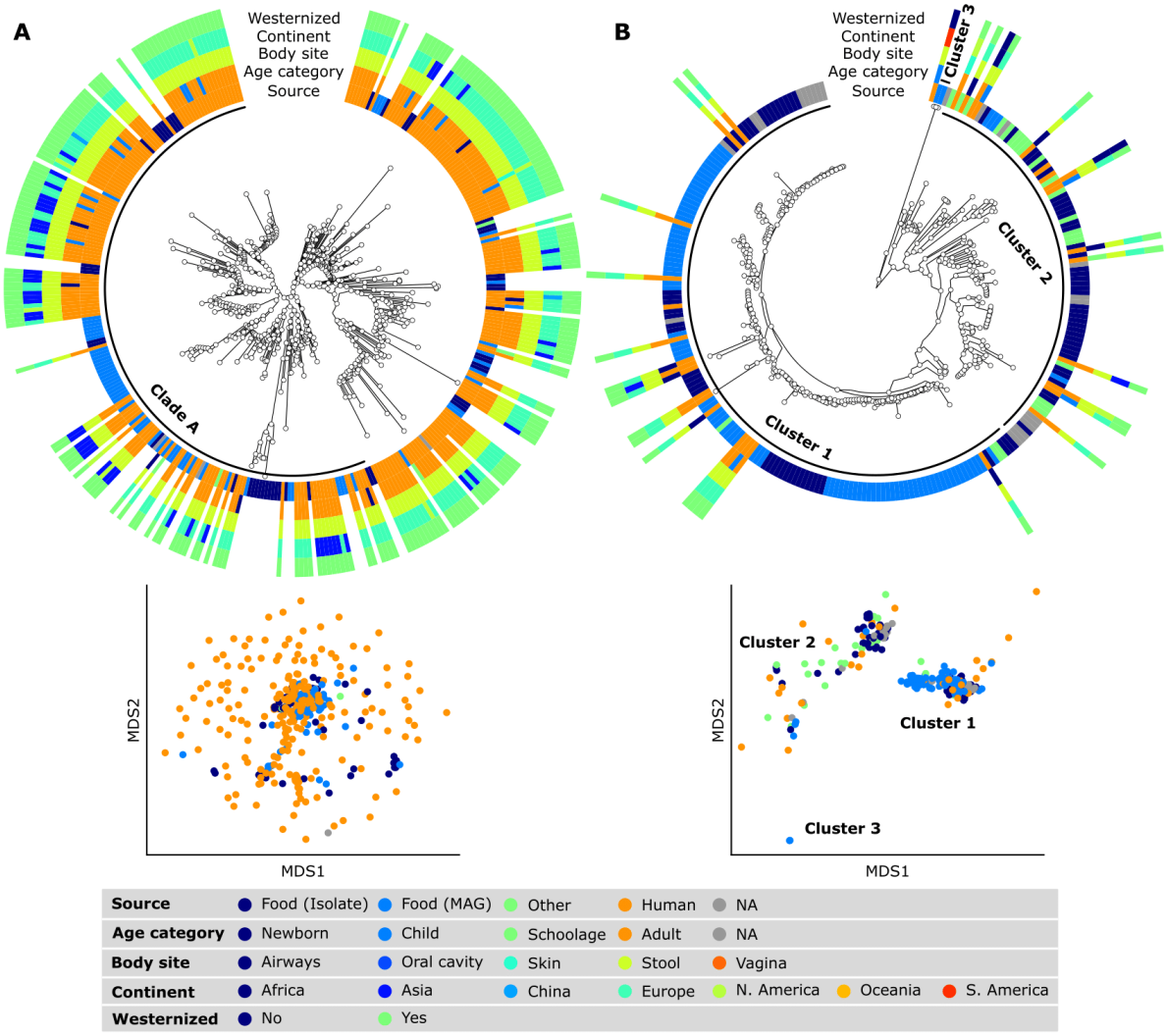


Figure 3.3. Comparative genomic analysis of the two most prevalent LAB identified in the human gut microbiome.

A) *S. thermophilus* is a genetically diverse species both in food and human sources with MAGs reconstructed from Asian gut metagenomes enriched in Clade A ($p < 1e\text{-}10$). **B)** *Lc. lactis subsp. lactis* is formed by three main clusters: Cluster 1 exhibits an overall low diversity and includes mostly food genomes related to cheese and dairy fermentation; Cluster 2 is dominated by environmental and raw vegetable products and more diverse human MAGs; Cluster 3 includes only two MAGs from nunu. Phylogenetic trees were built on species-specific marker genes and report five different metadata. Multidimensional scaling (MDS) on average nucleotide identity (ANI) distance is coloured with source information.

The SGB 7142 (N = 216, **Figure 3.4C**), labelled *Lb. casei/paracasei*, included reference genomes identified as both *Lb. casei* and *Lb. paracasei*, which, as recently highlighted, can be used interchangeably [94]. Within the combined species, we detected two main clusters, both of which occurred in food and human samples. The major cluster contained 86% of the available genomes, including all the dietary supplement strains and the majority (86%) of the

human MAGs. Consistent with its low abundance (**Figure 3.1**), only seven reference genomes and a single MAG were reconstructed from human samples for *Lb. helveticus* (**Figure 3.4D**). We identified three main subspecies, all occurring in both food and human sources. One cluster included all the dietary supplement strains, while genomes coming from food were predominantly spread across the other two groups.

Despite the high number of collected genomes (N = 369), *Lb. plantarum* was scarcely prevalent (1.8%) and abundant (av. 1.2%) in the gut (**Figure 3.1**), which was reflected by only 11 MAGs being reconstructed from human microbiomes (**Supplementary Figure 3.7**). All of these belonged to the main cluster (96% of the total genomes) associated with subsp. *plantarum*. A separate cluster was identified as subsp. *argentoratensis*, which was found in both food and human isolates but never reconstructed from metagenomes. The occurrence of multiple subspecies within the same SGB was also observed for eight additional LAB, i.e., *Lb. brevis*, *Lb. fermentum*, *Lb. johnsonii*, *Lb. reuteri*, *Lb. sakei*, *Leuconostoc lactis*, *Leuconostoc mesenteroides*, and *W. cibaria*, (**Supplementary Figure 3.7**). On the other hand, *Lc. garvieae* was spread into two different SGBs, with one comprising human MAGs from both westernized and non-westernized populations and the other only from non-westernized cohorts (**Supplementary Figure 3.7**). No genomes from food samples were collected at all for *Lb. crispatus*, *Lb. gasseri*, *Lb. jensenii*, *Lb. ruminis*, and *Lb. salivarius* (excluding a single isolate from ground beef). The non-food species *Lb. ruminis* and *Lb. salivarius* were quite prevalent in the gut with 145 and 42 MAGs reconstructed from human metagenomes, respectively (**Supplementary Figure 3.7**). For both species, isolate and MAGs extracted from the gut were distinct from genomes isolated from other animal microbiomes, which suggested long-term adaptation of these species to the human gut. We also identified a specific *Lb. salivarius* cluster associated with dietary supplement strains, which was found in a couple of saliva samples but never in the human gut.

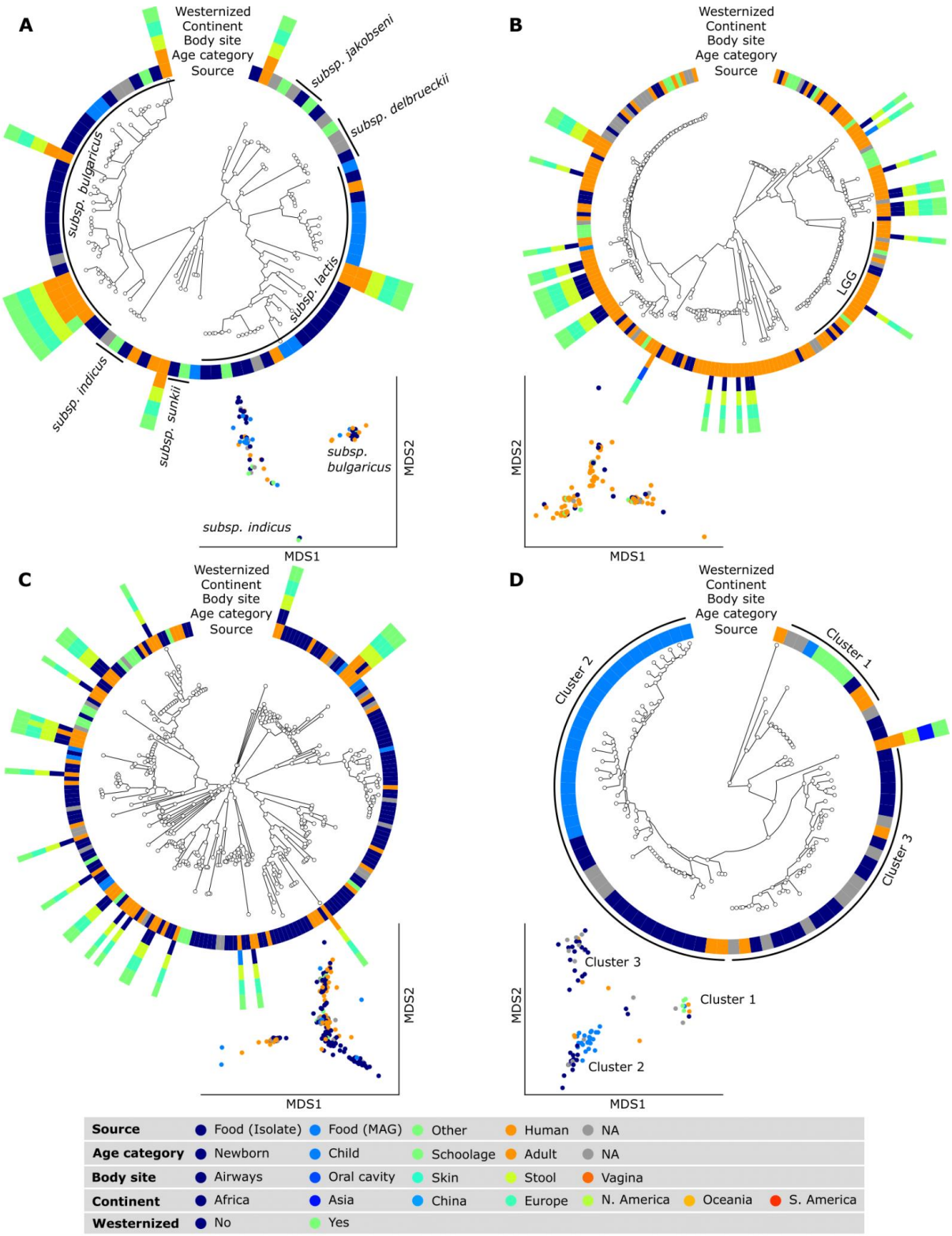


Figure 3.4. Comparative genomic analysis of relevant lactobacilli found in both food and human microbiomes.

A) *Lb. delbrueckii* is not prevalent in the gut, and the only two subspecies found in both food and human samples are subsps. *lactis* and. *bulgaricus*. Subsps. *delbrueckii*, *sunkii*, and *jakobsenii* are found in food, but never reconstructed from the gut. **B)** *Lb. rhamnosus* exhibits the greatest number of genomes from human isolates but is scarcely reconstructed from metagenomes. A specific cluster identifies the LGG strain. **C)** *Lb. casei/paracasei* includes reference genomes identified as both *Lb. casei* and *Lb.*

paracasei. We detect two main clusters both occurring in food and human samples. **D)** *Lb. helveticus* exhibits three main clusters, with Cluster 1 including all the dietary supplement strains (source in green), while food genomes are predominantly spread across the other two groups. Phylogenetic trees were built on species-specific marker genes and report five different metadata. Multidimensional scaling (MDS) on average nucleotide identity (ANI) distance is coloured with source information.

3.3.6 LAB occurrence in non-human primates is affected by captivity

We finally considered the set of 203 publicly available gut metagenomes from non-human primates (NHPs) that was recently retrieved, curated and processed with the same pipeline employed in this study [84]. It comprised 22 host species from 14 different countries in five continents. Among the 2,985 reconstructed MAGs, we found that only 46 of them (1.6%) were assigned to the Lactobacillales order (**Supplementary Table 3.8**), which suggested an overall low prevalence and abundance of LAB in the NHP gut microbiome. We found strong differences between MAGs retrieved from wild NHPs and those extracted from NHPs living in captivity. Wild NHPs generated 29 MAGs of LAB, with 66% of them associated with new species not available in public repositories and never found in human metagenomes, therefore likely representing bacteria peculiar to the NHP gut microbiomes. Ten MAGs were instead associated with kSGBs, with only five of them belonging to LAB species found also in human gut metagenomes such as *Lc. garvieae* (N = 3), *Lc. lactis*, and *Weissella cibaria*. Comparative genomics analysis highlighted that the strains harboured in NHPs were quite different from those reconstructed from human microbiomes (**Supplementary Figure 3.8**). Interestingly, the three MAGs of *Lc. garvieae* resembled more the strains found in non-westernized human populations in terms of nucleotide identity. No MAGs from lactobacilli were extracted at all from wild NHPs. A very different situation was observed in captive NHPs (**Supplementary Figure 3.8**), in which the 17 MAGs were exclusively reconstructed from kSGBs associated with multiple *Lactobacillus* species, i.e., *Lb. acidophilus*, *Lb. animalis* (N = 2), *Lb. johnsonii* (N = 4), *Lb. mucosae* (N = 2), *Lb. reuteri* (N = 5), and *Lb. salivarius* (N = 3). Strains of *Lb. reuteri* and *Lb. salivarius* found in NHPs were distinct from those extracted from human and food sources, which suggested possible host adaptation mechanisms. A stronger overlap among NHPs, human, and food MAGs was instead observed for the other species and likely linked to the sharing of strains due to the exposition of NHPs living in captivity to human-like environments and diets [95].

3.4 Discussion and Conclusions

In this chapter, we showed that food is likely the major source of LAB in the human gut microbiome. This was accomplished by conducting a large-scale meta-analysis that integrated taxonomic profiling and comparative genomics from almost ten thousand metagenomes from human and food sources in addition to reference genomes from public repositories. We focused the analysis on the thirty LAB that exhibited a prevalence greater than 0.1% in the human gut, which resulted mainly in species of potential food origin, including LAB occurring in probiotic supplements, in addition to non-food origin species such as *Lb. mucosae*, *Lb. ruminis*, and *Lb. salivarius*. The comparative genomics suggested that closely related strains are present in both food and gut microbiome. While such evidence does not exclude the possibility of other potential sources of LAB strains in the gut, we believe that being fermented foods the principal ecological niche for LAB in nature, our results support the hypothesis that food is the major source of LAB for the gut microbiome. While we considered the currently available taxonomic nomenclature, a substantial reclassification of the genus *Lactobacillus* into 25 novel genera enclosing the current *Lactobacillus* species was recently proposed [96]. The new *Lactobacillus* genus incorporates only the species included in the *Lb. delbrueckii* group.

We found an overall limited amount of LAB in the gut in terms of prevalence and relative abundance, however several species exhibited non-negligible contributions that deserve attention for potential probiotic potentials. There was no evident correlation between prevalence and relative abundance of the different LAB species in the human samples. The most prevalent LAB species was *S. thermophilus*. Its role as a gut microbiome member is questioned. However, the mechanisms and metabolic features that lead to it being regarded as a candidate probiotic species have been studied and debated, especially in terms of resistance to gastrointestinal barriers and potential positive health effects [97]. Beyond being one of the two LAB widely employed for yogurt making, *S. thermophilus* is also employed as starter cultures for many cheeses characterized by a thermophilic fermentation. Continuous exposure to *S. thermophilus* through cheese and yogurt consumption can be a likely explanation of its prevalence in human gut samples as resulted in this study.

We detected a remarkable prevalence in the gut also for *Lc. lactis*, which is widespread in cheeses produced by mesophilic fermentation. Albeit recognized as a transient member of the gut community, higher levels of this species were found in buttermilk consumers [98]. In addition, strains of *L. lactis* have been shown to survive the gastrointestinal stress and this

species can be considered to potentially convey health benefits by antimicrobial activity through bacteriocin production against clostridia, to boost the immune system, and to be potentially used as a vehicle of interesting beneficial properties such as antimicrobial activity [99,100].

The prevalence of LAB in the human gut was strongly affected by lifestyle [101], intended here as possible consumption of fermented foods that are characteristics of specific geographical regions. Unfortunately, direct associations of genomic data with dietary patterns could not be achieved as dietary records documenting systematic food consumption in the human public cohorts considered were not available. Minor associations between gut microbiota and consumption of plant fermented foods were very recently found within the American gut cohort. A few LAB species were linked to fermented plant food consumers and included *Lb. acidophilus*, *Lb. brevis*, *Lb. kefiranofaciens*, *Lb. parabuchneri*, *Lb. helveticus* and *Lb. sakei*. Interestingly, the authors highlighted that the stool detection of LAB may be a useful tool to verify the reliability of self-reported dietary information on fermented foods consumption [102].

In our study, LAB species widely occurring in dairy products and yogurt, such as *S. thermophilus* and lactobacilli, were more prevalent in westernized populations, while the heterofermentative *Leuconostoc* and *Weissella*, likely carried as part of the epiphytic microbiota of raw vegetables [79], fermented vegetables [103], and cereal-based fermented foods [104] were more common in the non-westernized cohorts. We could speculate that this pattern was linked to the habitual consumption of foods and diets that were characteristics of the specific geographical areas. For example, non-westernized populations that have a higher consumption of raw plants and plant based fermented foods were enriched in heterofermentative cocci LAB, while the very low prevalence of *Lc. lactis* and *S. thermophilus* in multiple Chinese cohorts reflects the low consumption of dairy products by the Chinese population [105].

We conducted an extensive comparative genomic analysis by integrating reference genomes and MAGs from human, food, and environmental sources. This opportunity was previously prevented even for prevalent species such as *S. thermophilus* and *Lc. lactis* due to the lack of reference genomes acquired from human sources in public repositories. We identified a general overlap among genomes from food and gut sources, which suggested again food as the main source of LAB in the human gut. To this end, we conducted a preliminary analysis devoted to

evaluate potential differences in functions of strains between food and gut sources, that we limited to *Lc. lactis* and *S. thermophilus* due to their large number of MAGs reconstructed in this study (see **Methods**). We found 266 (247 in food) and 323 (275 in food) differently prevalent genes ($p < 0.05$) for *Lc. lactis* and *S. thermophilus*, respectively, after removing genes encoding for unidentified functions or occurring redundantly in both food and gut groups (differently prevalent sugar metabolism genes are listed in **Supplementary Table 3.9**). However, such differences did not suggest remarkable potential functional differences between food and gut genomes, which was consistent with the comparative genomics and phylogenetic results shown in **Figure 3.3**. At the same time, we identified an increase of unannotated genes in the gut genomes for both species, which agreed with the scarcity of reference genomes from human sources in public repositories. This may reflect further differences of strains found in the human gut that are currently unexplored due to the incompleteness of available functional databases [106]. Functional differences may suggest a possible adaptation of the food LAB to the gut environment. However, such mechanisms of adaptations cannot occur in strains that are part of a transient microbiome and would only take place for those LAB that more stably colonize the gut environment. This opens the need to conduct new analyses focused on the isolation of these microorganisms from the gut and their more in-depth functional characterization, also based on phenotypic traits. Different patterns were observed for typical non-food origin species such as *Lb. ruminis* and *Lb. salivarius*. By comparing human genomes with those found in other environments including animal microbiomes, we identified a strong adaptation of these species to the human gut, which suggested that these species are more specific and persistent for the human host (**Supplementary Figure 3.7**).

Some of the analysed LAB exhibited distinct groups with human and food genomes clustering together, which indicated the presence in the gut of different strains potentially coming from different food sources. For example, the genomes of *Lb. delbrueckii* reconstructed from the gut appeared to cluster in two main groups associated with subsp. *bulgaricus* and subsp. *lactis*, which were representative of LAB in yogurt and cheese, respectively. Multiple subclusters were identified also in *Lb. rhamnosus*, with only 17% of the reconstructed human MAGs corresponding to the strain GG largely used in probiotic supplements. These species, along with others such as *Lb. casei*, *Lb. plantarum,* and *Lb. reuteri* have been largely explored due to their probiotic potential. However, their general low prevalence and abundance in the human gut suggested that they are unlikely to be long-term residents of the gut microbiota. However, we used only faecal samples as representative of the gut microbiome, while such species may

be more tightly adhered to the gut epithelium and therefore less detectable in stool specimens [107].

Finally, we highlight the importance of considering computational approaches such as those exploited in this chapter. Strain-level genome comparison is fundamental to track the resilience and persistence of probiotic LAB in the human gut and can be a useful approach to be adopted in clinical trials aimed at evaluating the efficacy of microbial strains for gut health. Additionally, the same methodologies can be considered to evaluate the prevalence and resilience of non-food microorganisms that are currently studied as candidates for next generation probiotics. Such knowledge and approaches can be useful for an informed design of functional foods, conveying health benefits upon daily consumption beyond their nutritional value. Several functional foods are enriched with probiotic microbial strains and their fate in pre-clinical and clinical trials can be efficiently and reliably monitored by culture-independent genome reconstruction and comparison to help assess both their efficacy as probiotics and the quality of the functional food.

The interest in LAB will keep the scientific community active in studies of their genomics and evolution. Some of the LAB species occurring in the gut can surely arise from the consumption of fermented foods or probiotic preparations. However, efforts in research and isolation of LAB from human specimens would be desirable in the future in order to have further evidence on their specific genomic features that may better reflect adaptation to the complex gut ecosystem.

5 Conclusions

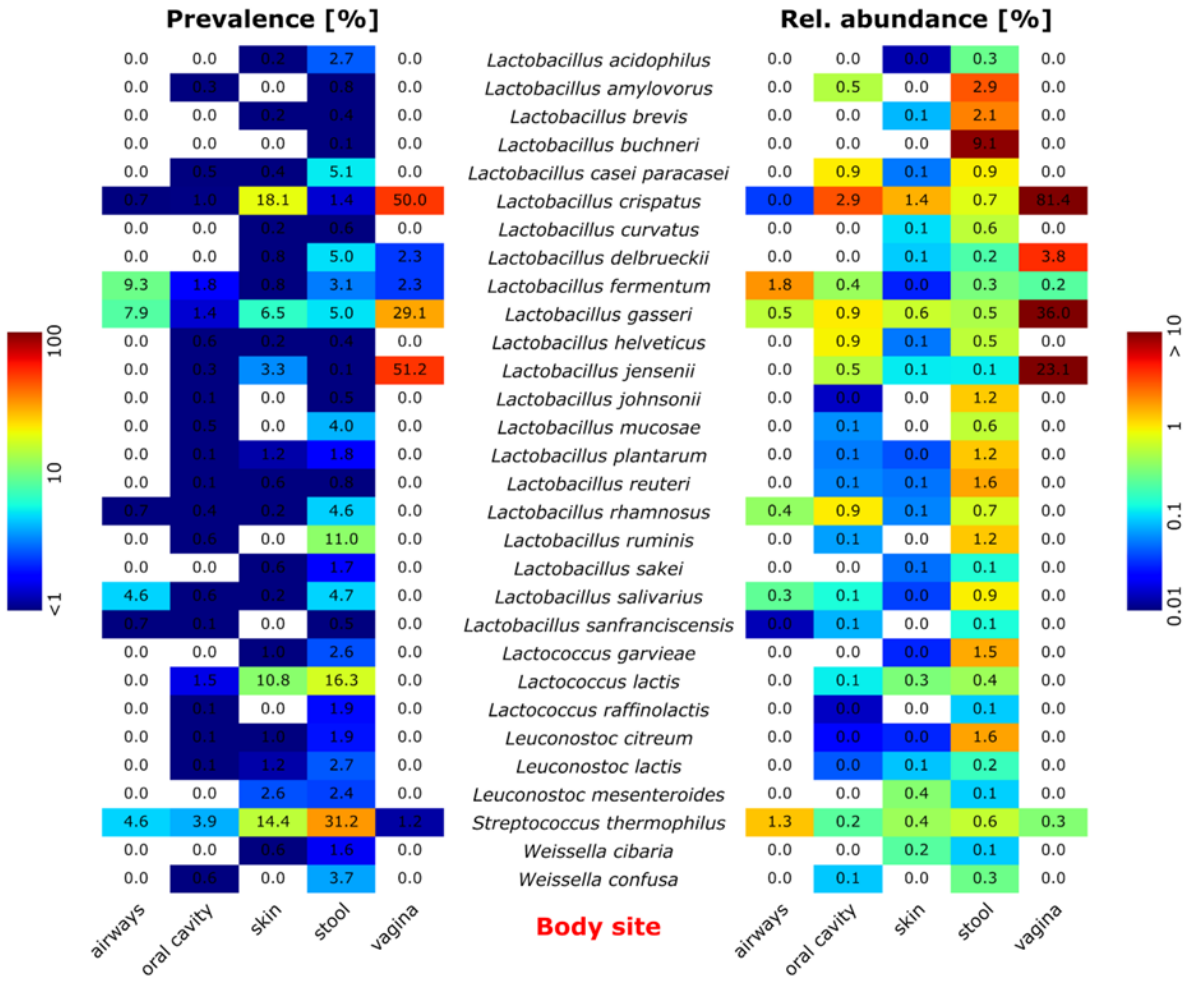
In my doctoral research we have implemented and validated a computational methodology aiming at estimating the number of sub-species in selected microbial species belonging to the human and food microbiomes. This is based on the analysis of reference genomes and MAGs and the use of advanced clustering techniques. The methodology has been mainly validated on synthetic data and also applied in real scenarios by considering two microbial families of great relevance in the food science field as LAB and *Bifidobacteriaceae*.

We conducted a large-scale analysis of LAB from (meta-)genomics data in which we considered 9,445 metagenomes from human samples and 303 food metagenomes. We demonstrated that the prevalence and abundance of LAB species in stool samples using state-of-the-art methodology is generally low and linked to age, lifestyle and geography, with *Streptococcus thermophilus* and *Lactococcus lactis* being most prevalent. We also identified genome-based differences between food and gut microbes. Overall, the large-scale genome-wide analysis demonstrated that closely related LAB strains occur in both food and gut environments and provides unprecedented evidence that fermented foods can be indeed regarded as possible sources of LAB for the gut microbiome.

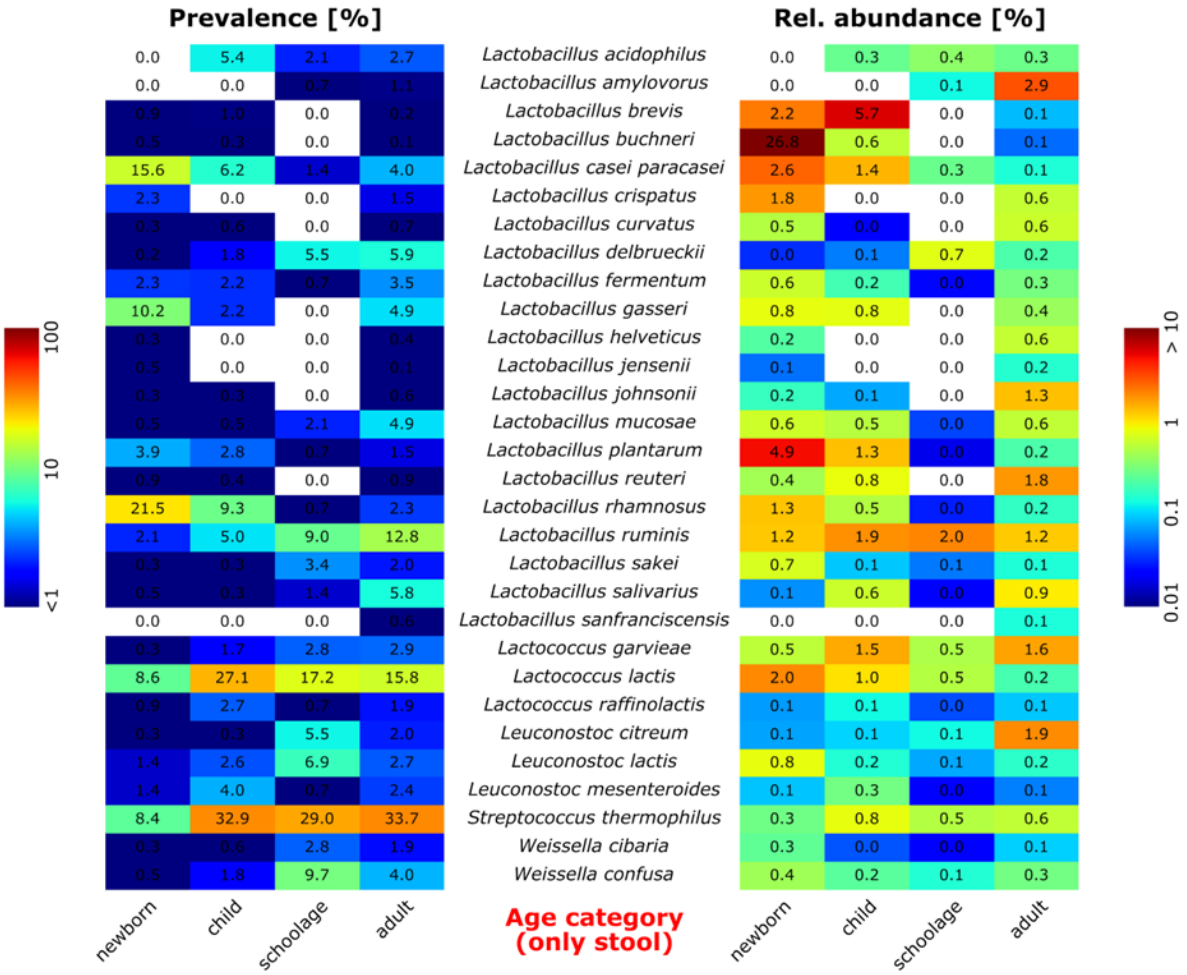
We also performed a large-scale analysis for species belonging to the *Bifidobacteriaceae* family. Similarly to the analysis conducted for LAB, we demonstrated that their prevalence and abundance is linked to age, lifestyle and geography. We found variable predictive capabilities in estimating host phenotypes from microbiome data using *Bifidobacteriaceae* relative abundances as the only information. Moreover, we identified multiple sub-species in different prevalent Bifidobacteriaceae species which are new with respect to what reported in the literature.

We established the methodological framework to detect and characterize sub-species in food and human microbiomes. Although our methodology showed high performances, it could be improved by validating and testing more scenarios by varying setting parameters in synthetic data generation (i.e., number of clusters, probability distribution, number of observations) and computed distance among observations (i.e. euclidean). While more in depth analysis was performed on LAB and *Bifidobacteriaceae* similar efforts may be performed on other species of relevance for the human microbiome and potentially linked to human health.

6 Supporting informations

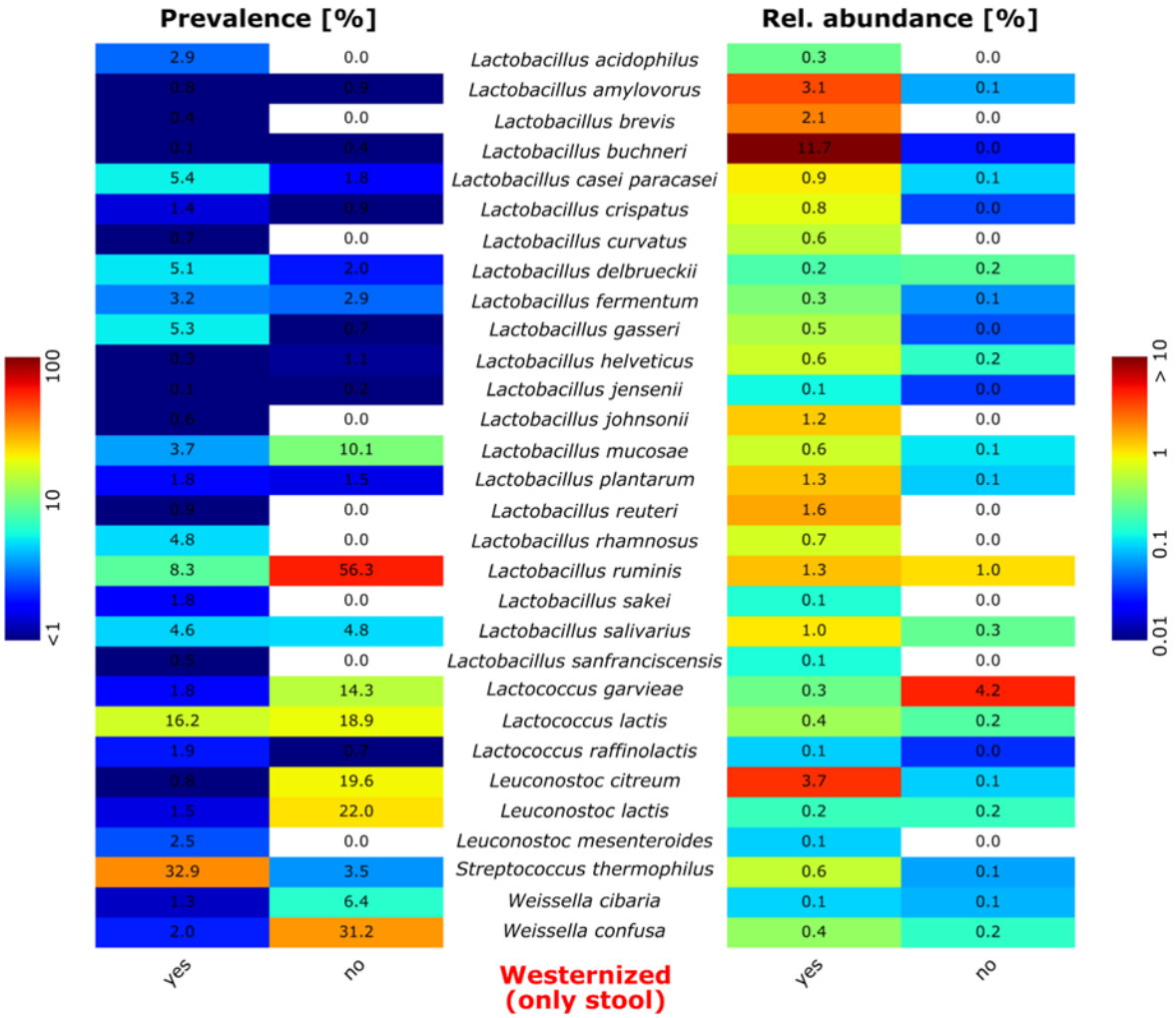


Supplementary Figure 3.1. Average prevalence and relative abundance of LAB species from the human microbiome stratified by body site.
Average relative abundance is computed on positive samples only. Raw data in **Supplementary Table 3.2**.

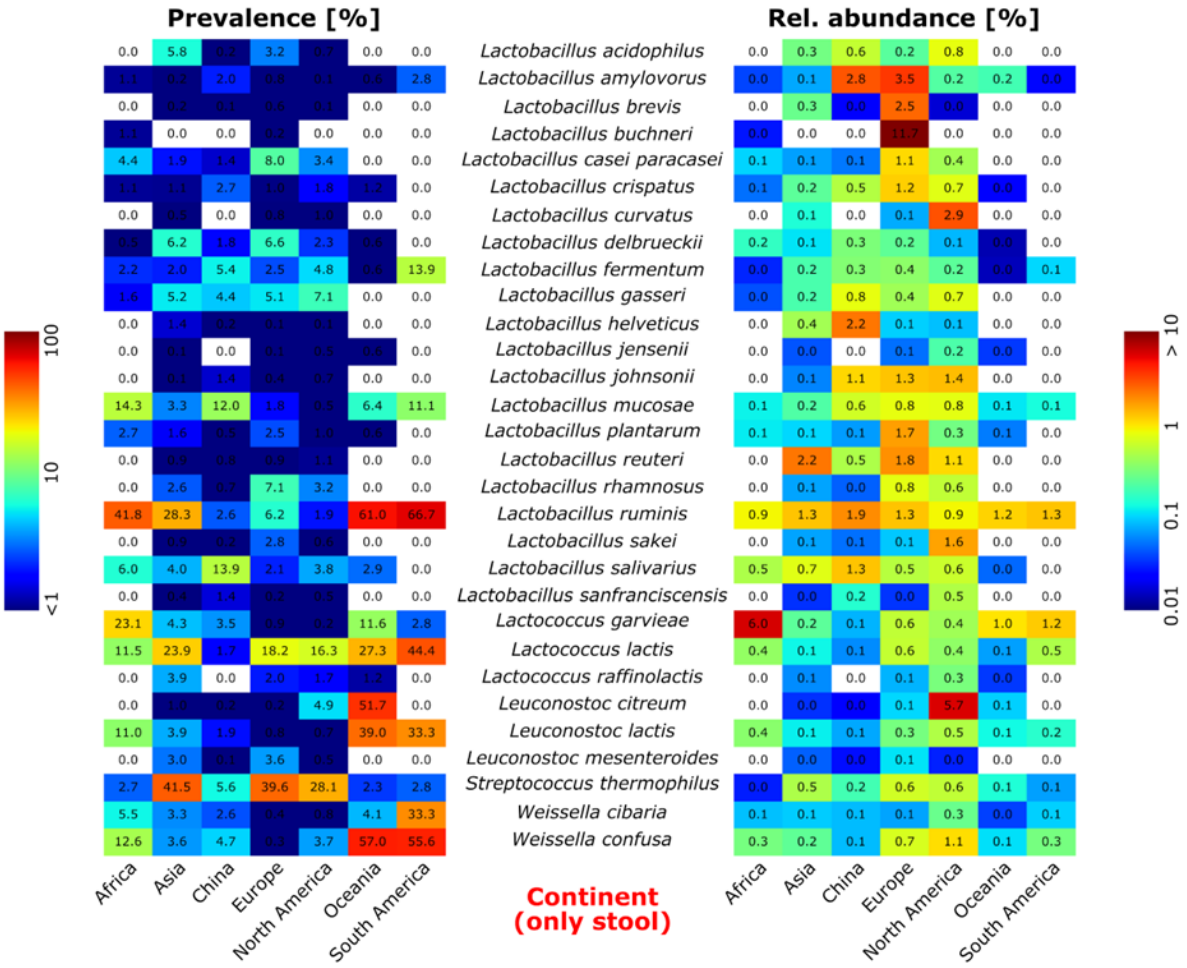


Supplementary Figure 3.2. Average prevalence and relative abundance of LAB species from the human microbiome stratified by age category.

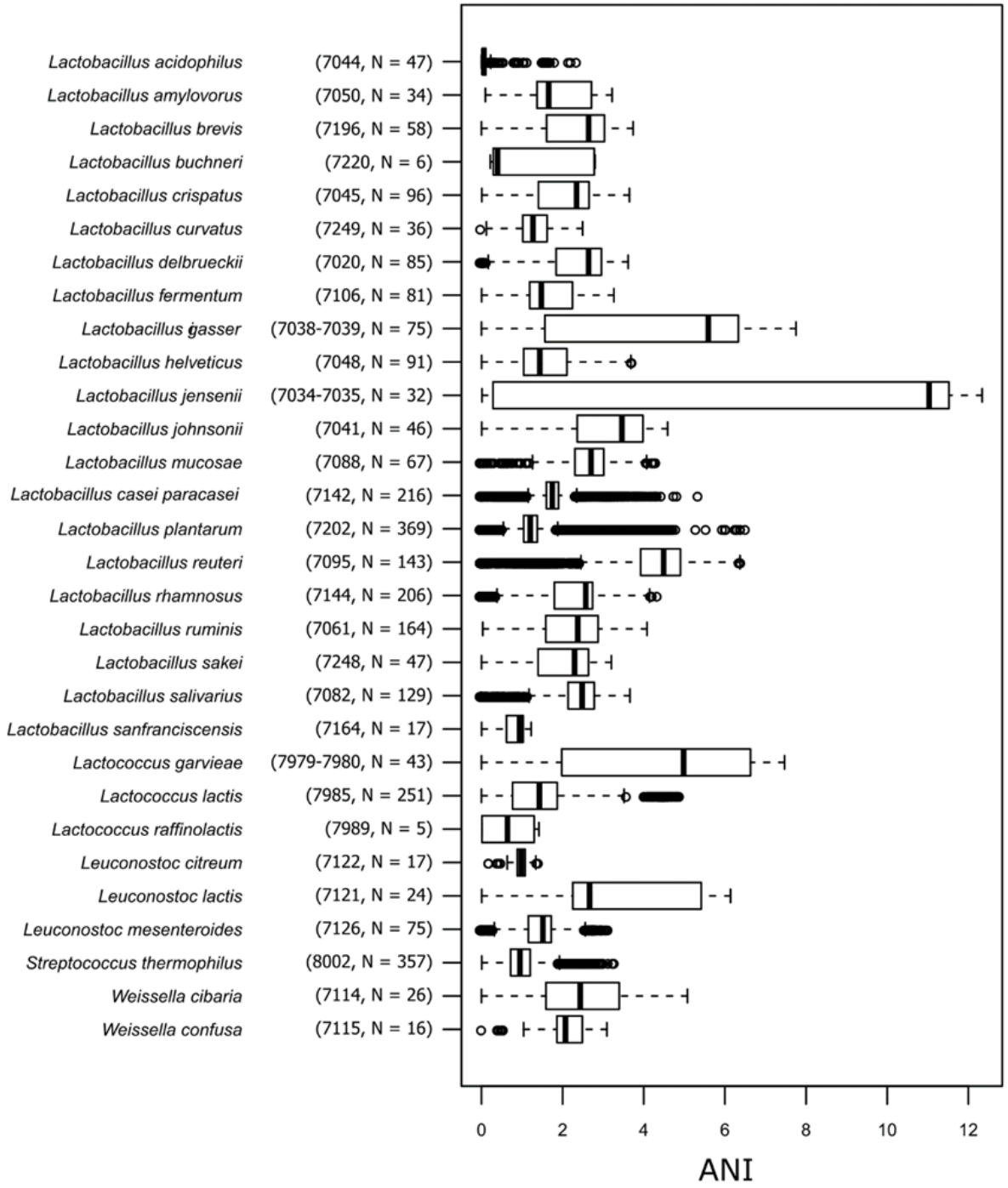
Statistics refer to stool samples only. Average relative abundance is computed on positive samples only. Raw data in **Supplementary Table 3.2**.



Supplementary Figure 3.3. Average prevalence and relative abundance of LAB species from the human microbiome stratified by westernized lifestyle.
Statistics refer to stool samples only. Average relative abundance is computed on positive samples only. Raw data in **Supplementary Table 3.2**.

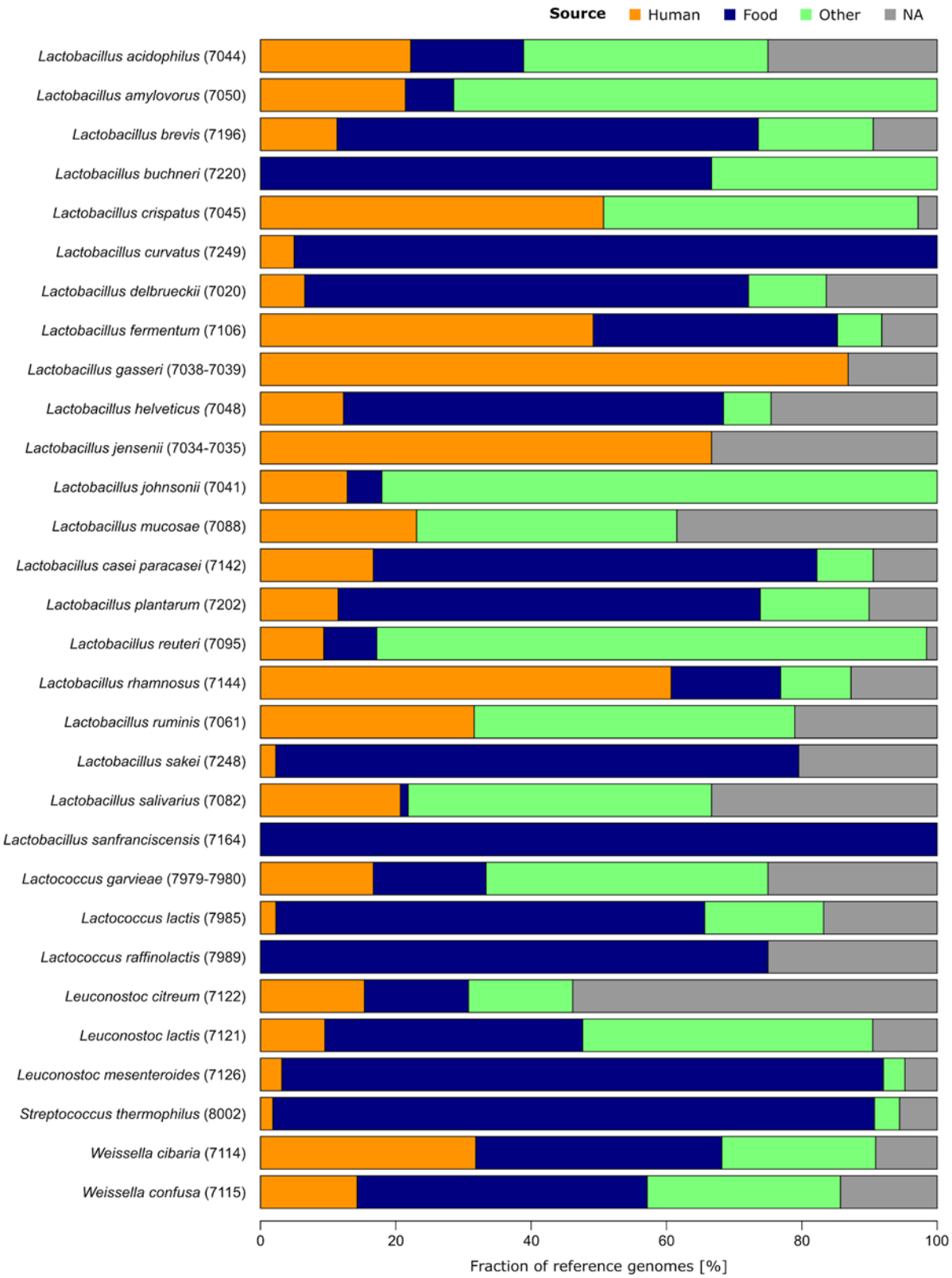


Supplementary Figure 3.4. Average prevalence and relative abundance of LAB species from the human microbiome stratified by continent.
Statistics refer to stool samples only. Average relative abundance is computed on positive samples only. Raw data in **Supplementary Table 3.2**.



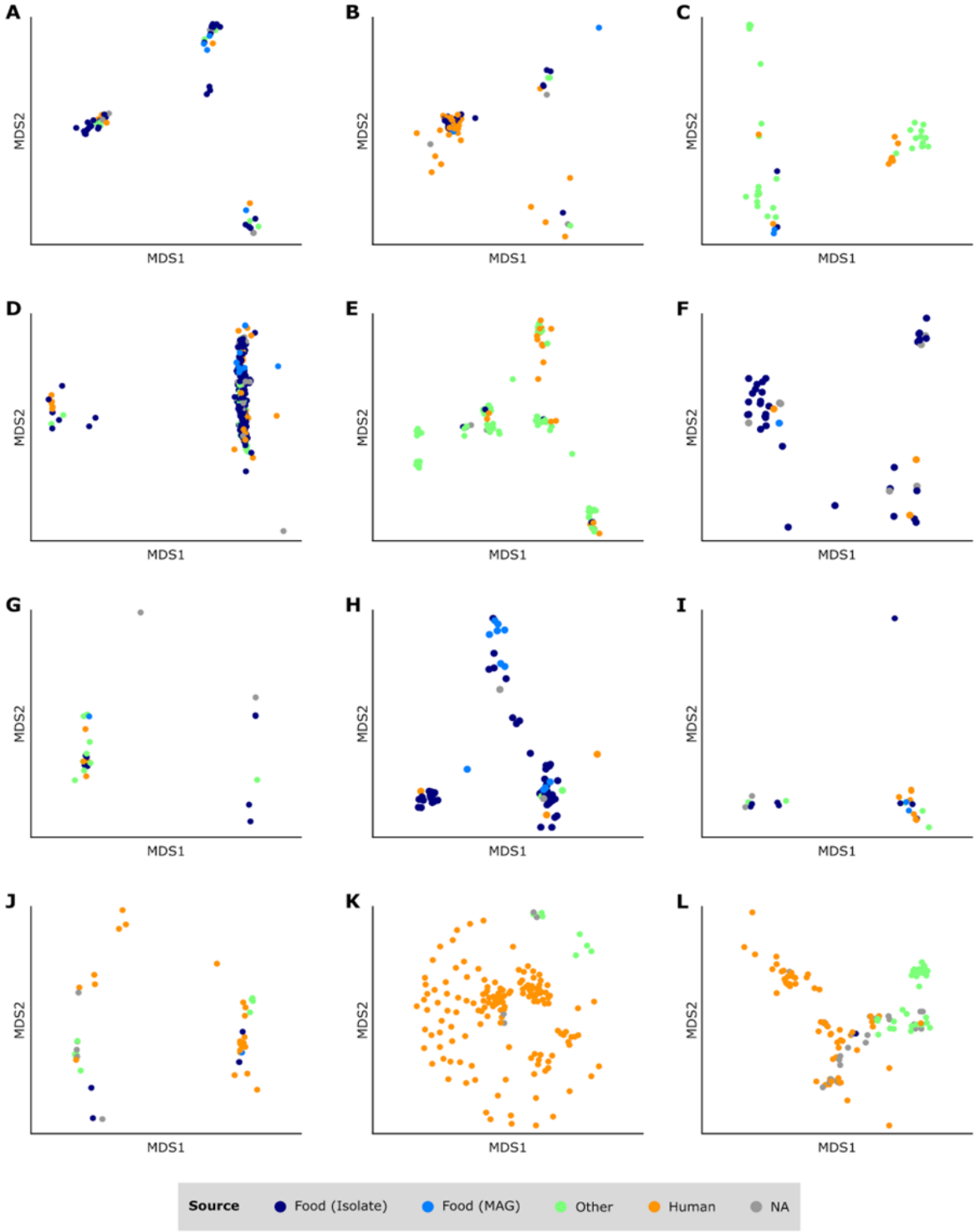
Supplementary Figure 3.5. Average nucleotide identity (ANI) for the 30 selected LAB species on the set of reference genomes and MAGs.

The ANI is computed using FastANI and excluding genomes having completeness < 80%. Numbers in parenthesis represent the SGB ID and the sample size. Three species (i.e., *Lb. gasseri*, *Lb. jensenii*, and *L. garvieae*) span two SGBs and are identified with two numeric identifiers. The lower and upper hinges correspond to the first and third quartiles (i.e., the 25th and 75th percentiles). The line inside the box represents the median value. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge (where IQR is the inter-quartile range, or distance between the first and third quartiles). The lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge. Points beyond the end of the whiskers are outliers and plotted individually.



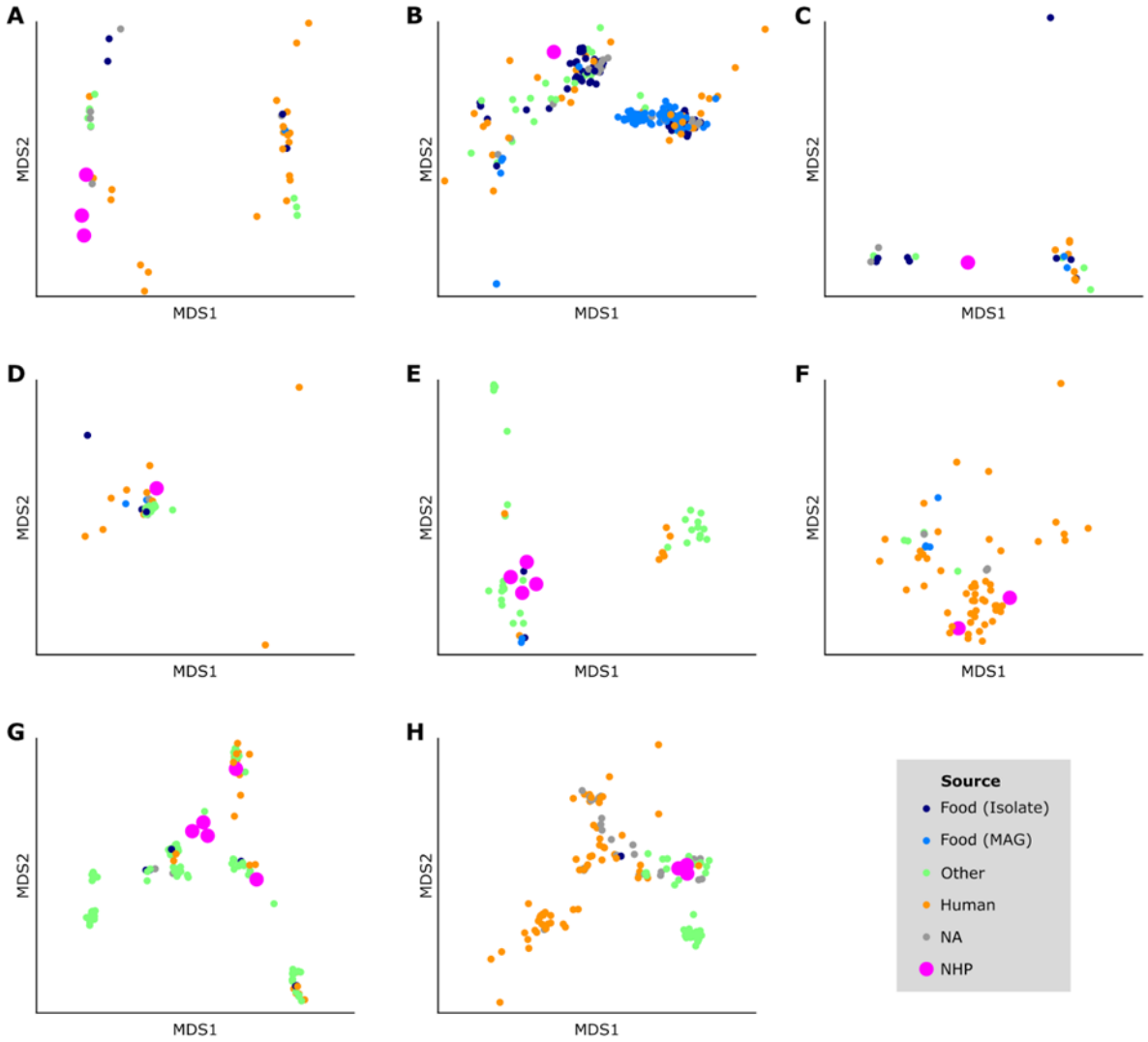
Supplementary Figure 3.6. Fraction of reference genomes per source type for the 30 selected LAB species.

The same plot grouped at genus-level is reported in **Figure 3.2C**.

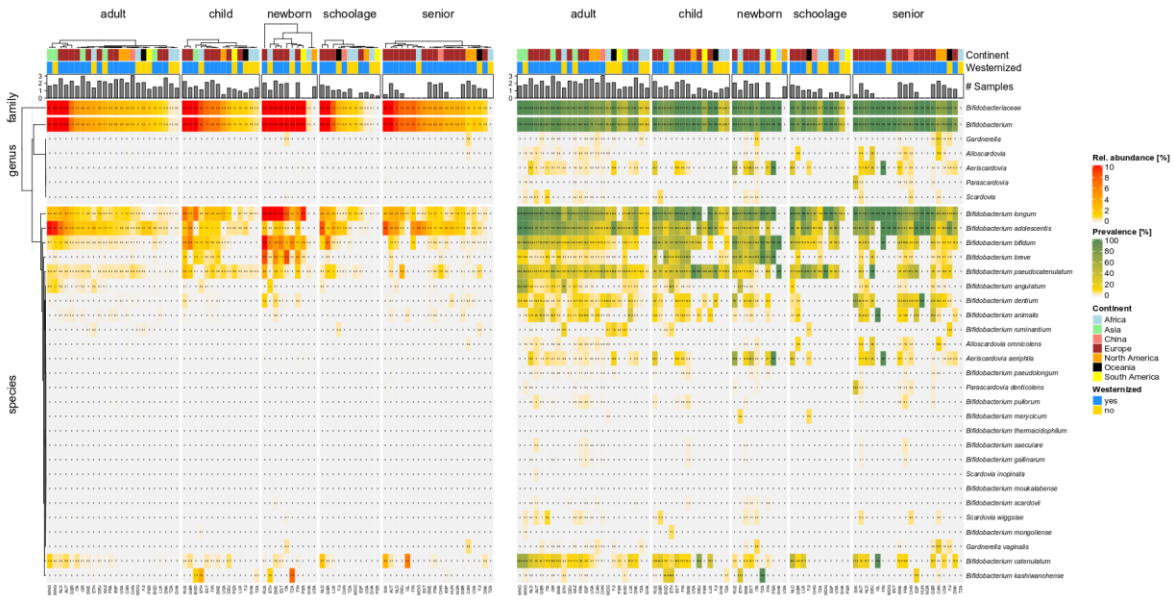


Supplementary Figure 3.7. Comparative genomic analysis of relevant LAB species.
Results refer to **A)** *Lb. brevis* (N = 58); **B)** *Lb. fermentum* (N = 81); **C)** *Lb. johnsonii* (N = 46); **D)** *Lb. plantarum* (N = 369); **E)** *Lb. reuteri* (N = 143); **F)** *Lb. sakei* (N = 47); **G)** *Leuconostoc lactis* (N = 24); **H)** *Leuconostoc mesenteroides* (N = 75); **I)** *W. cibaria* (N = 26); **J)** *L. garvieae* (N = 43); **K)** *Lb. ruminis* (N = 164); and **L)** *Lb. salivarius* (N = 129). **A)-I)** Species having occurrence of multiple subspecies into the same SGB with genomes from both food and human gut sources; **J)** *L. garvieae* is spread into two

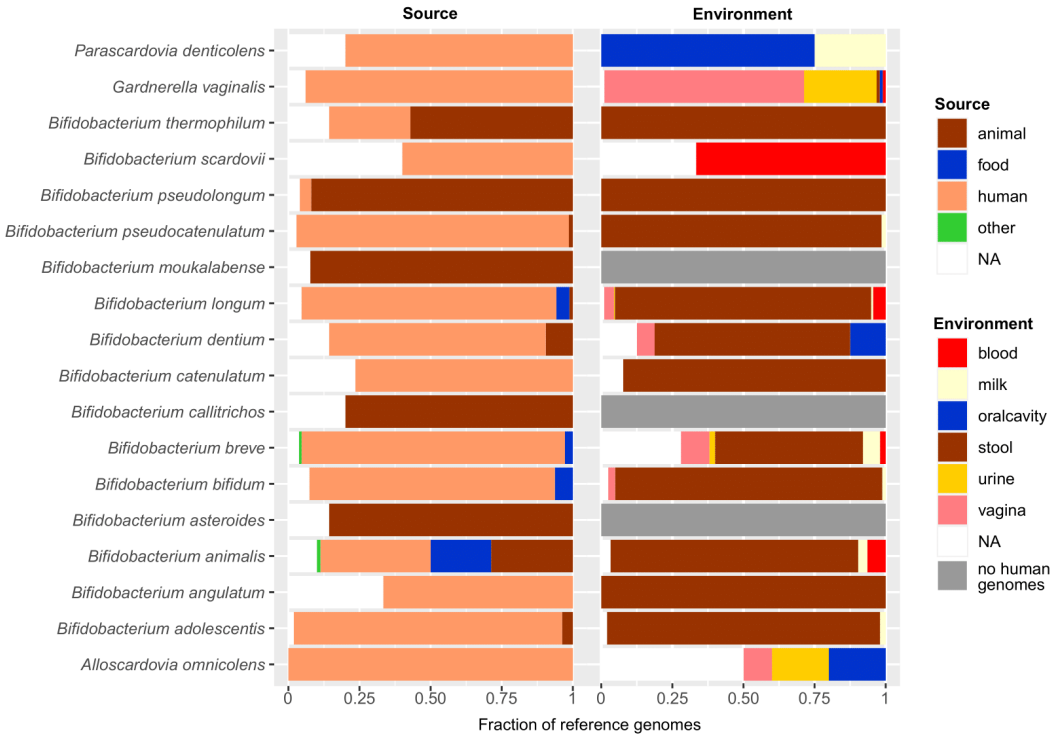
different SGBs; and **K)-L)** the typical non-food origin species *Lb. ruminis* and *Lb. salivarius* exhibit genomes extracted from the gut that are distinct from genomes isolated from other environments and animal microbiomes, suggesting long-term adaptation of these species to the human gut. Multidimensional scaling (MDS) on average nucleotide identity (ANI) distance is colored with source information. Plots for additional species are reported in **Figure 3.3** and **Figure 3.4**.



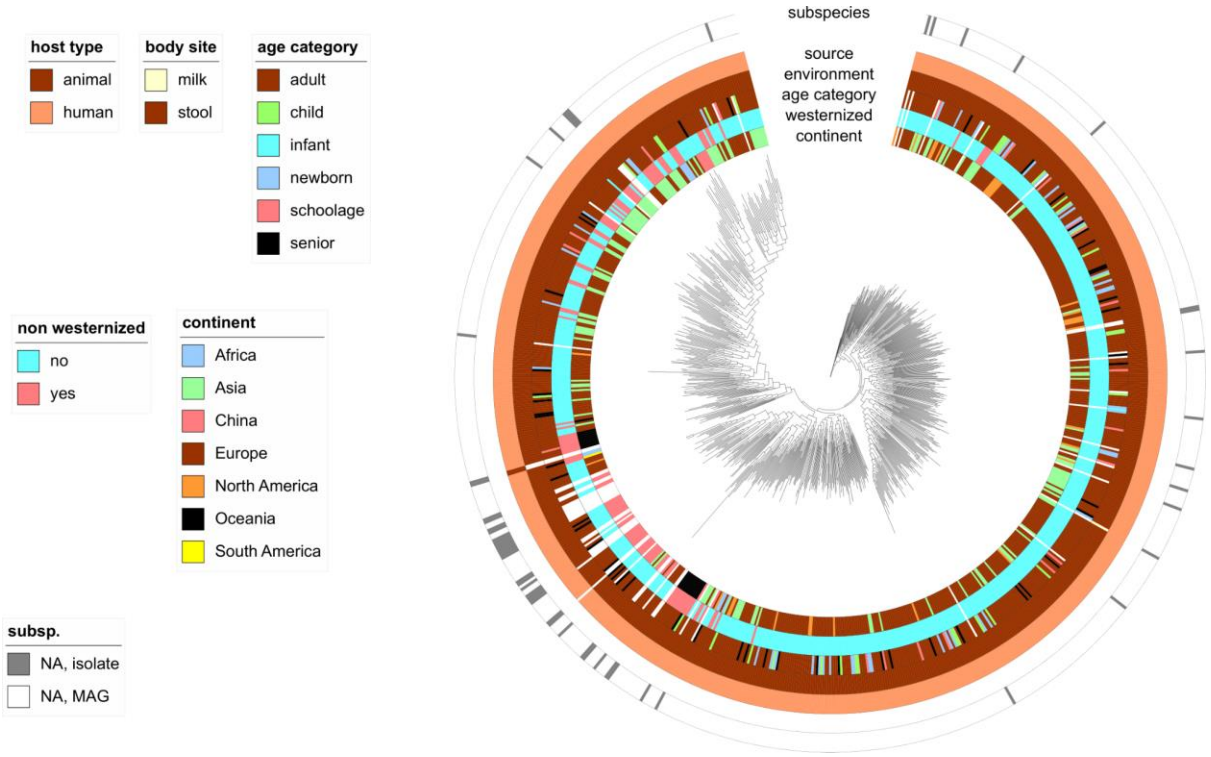
Supplementary Figure 3.8. Comparative genomic analysis of the LAB species reconstructed from NHP metagenomes that overlap with MAGs extracted from the human gut.
Results refer to **A)** *Lc. garvieae* (N = 3 NHP MAGs); **B)** *Lc. lactis* (N = 1); **C)** *Weissella cibaria* (N = 1); **D)** *Lb. acidophilus* (N = 1); **E)** *Lb. johnsonii* (N = 4); **F)** *Lb. mucosae* (N = 2); **G)** *Lb. reuteri* (N = 5); and **H)** *Lb. salivarius* (N = 3). NHP MAGs were retrieved from **A)-C)** wild NHPs and **D)-H)** NHPs living in captivity. Multidimensional scaling (MDS) on average nucleotide identity (ANI) distance is coloured with source information.



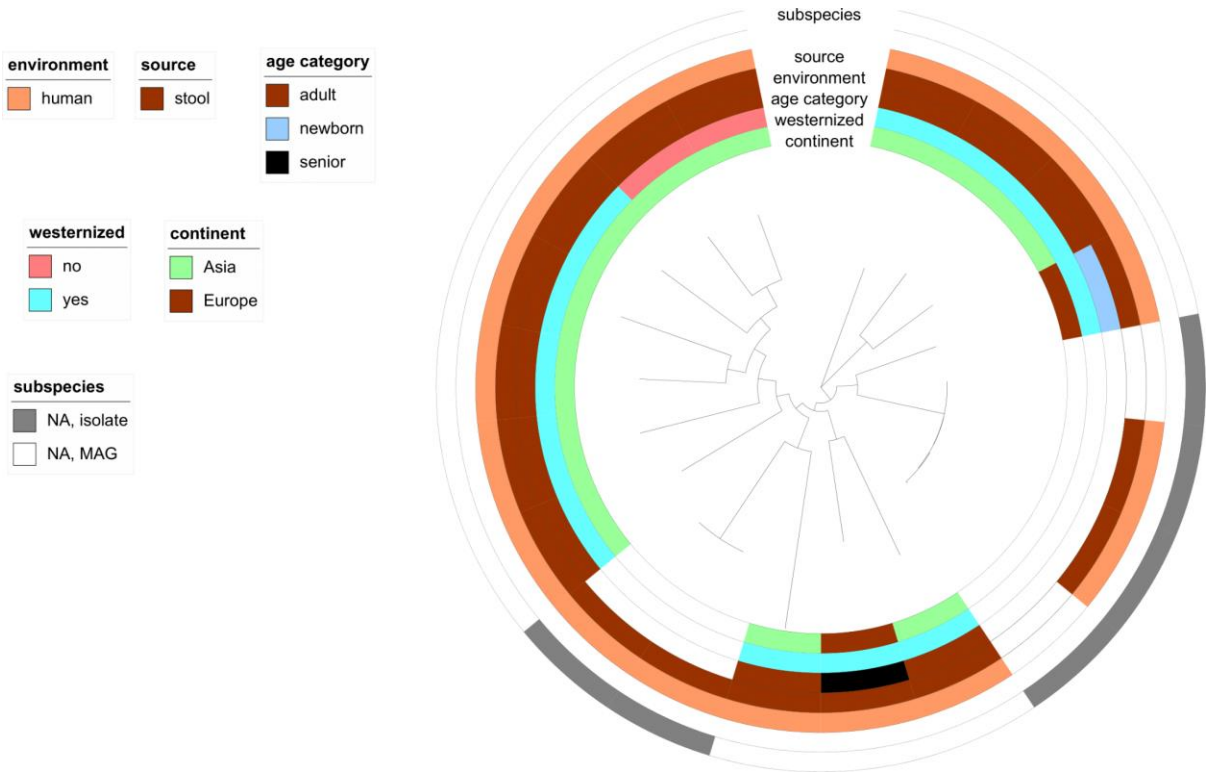
Supplementary Figure 4.1. Average relative abundance and prevalence of Bifidobacteriaceae spp. in the human gut microbiome across multiple host phenotypes.
Values are stratified by body site, age category, lifestyle, and geography.



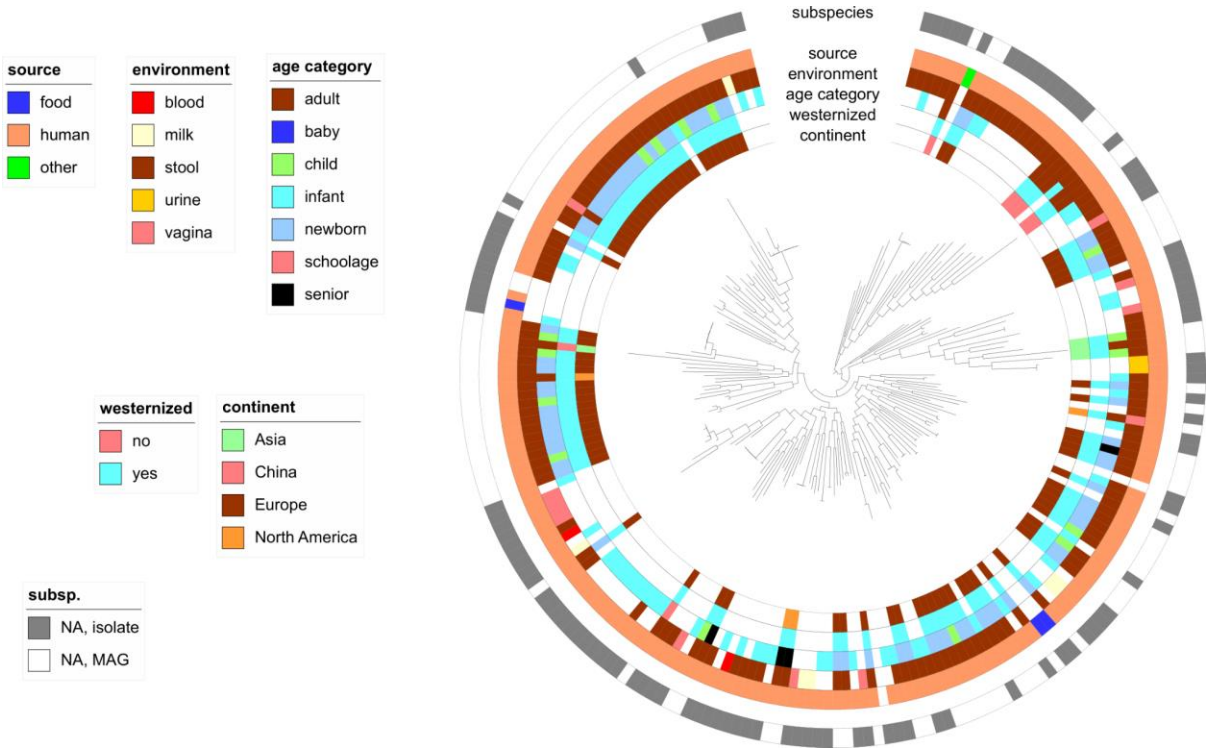
Supplementary Figure 4.2. Fraction of isolate genomes per source and environment type considered in this study.
Only the *Bifidobacteriaceae* species having at least 5 genomes are reported.



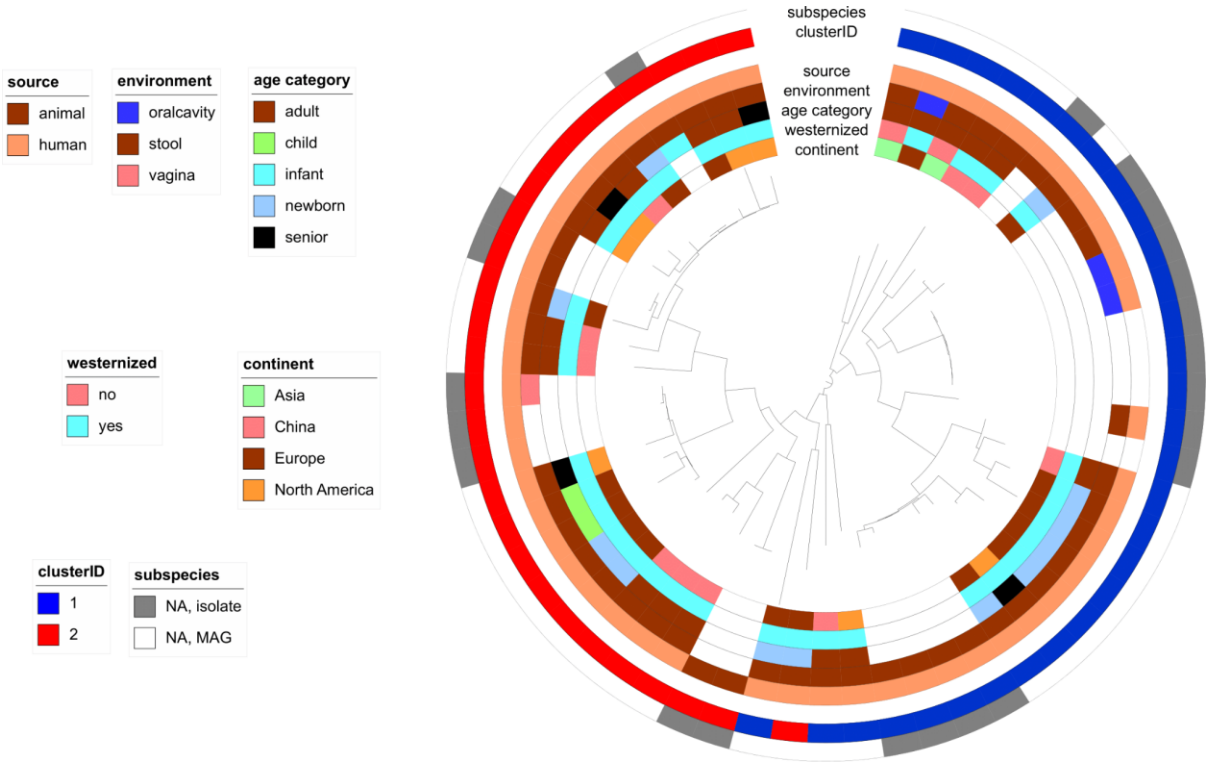
Supplementary Figure 4.3. Phylogenetic analysis for *B. adolescentis* by integrating isolate genomes and MAGs. The phylogenetic tree reports multiple host characteristics.



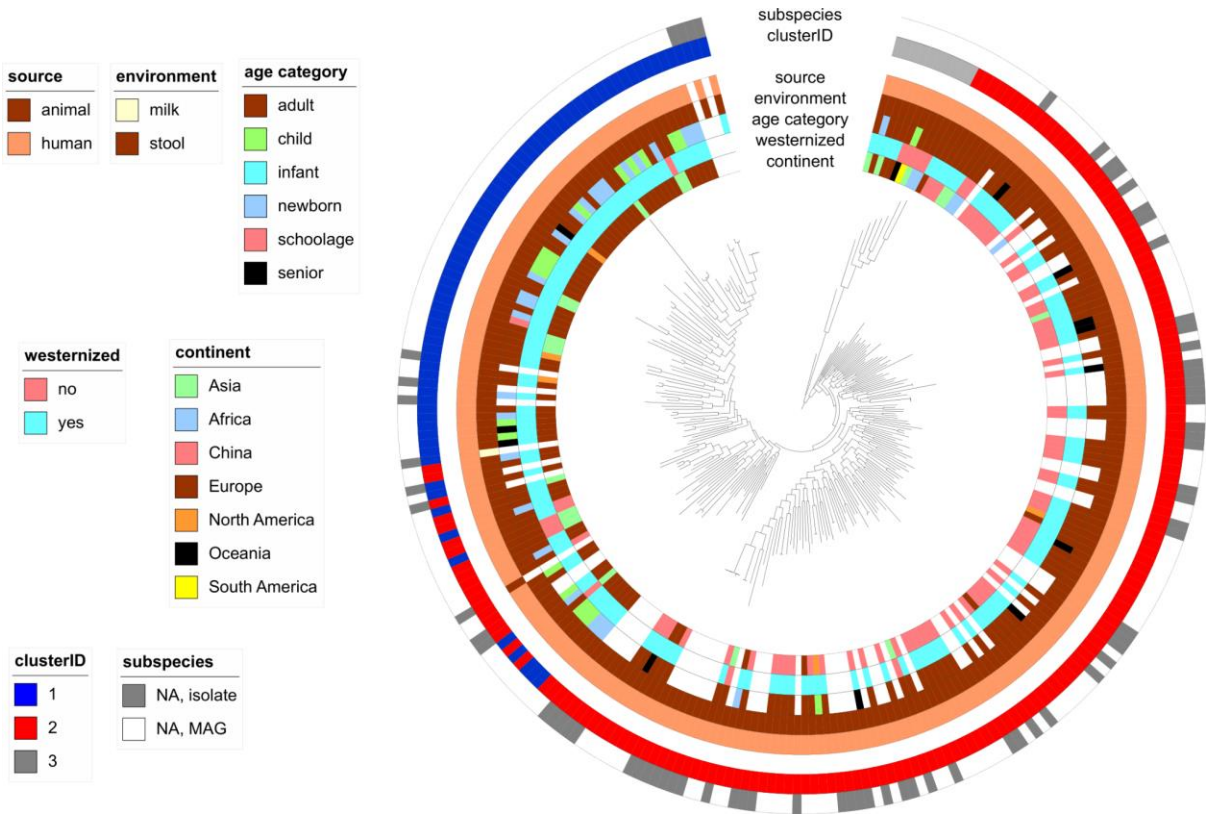
Supplementary Figure 4.4. Phylogenetic analysis for *B. angulatum* by integrating isolate genomes and MAGs. The phylogenetic tree reports multiple host characteristics.



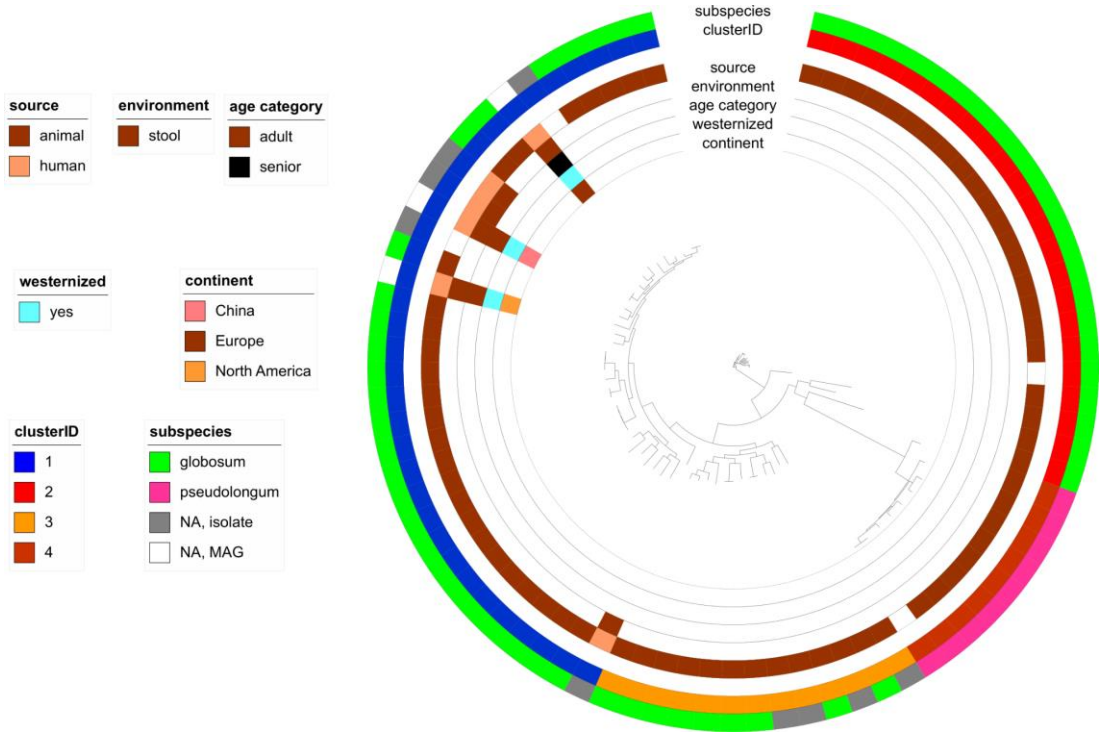
Supplementary Figure 4.5. Phylogenetic analysis for *B. breve* by integrating isolate genomes and MAGs. The phylogenetic tree reports multiple host characteristics.



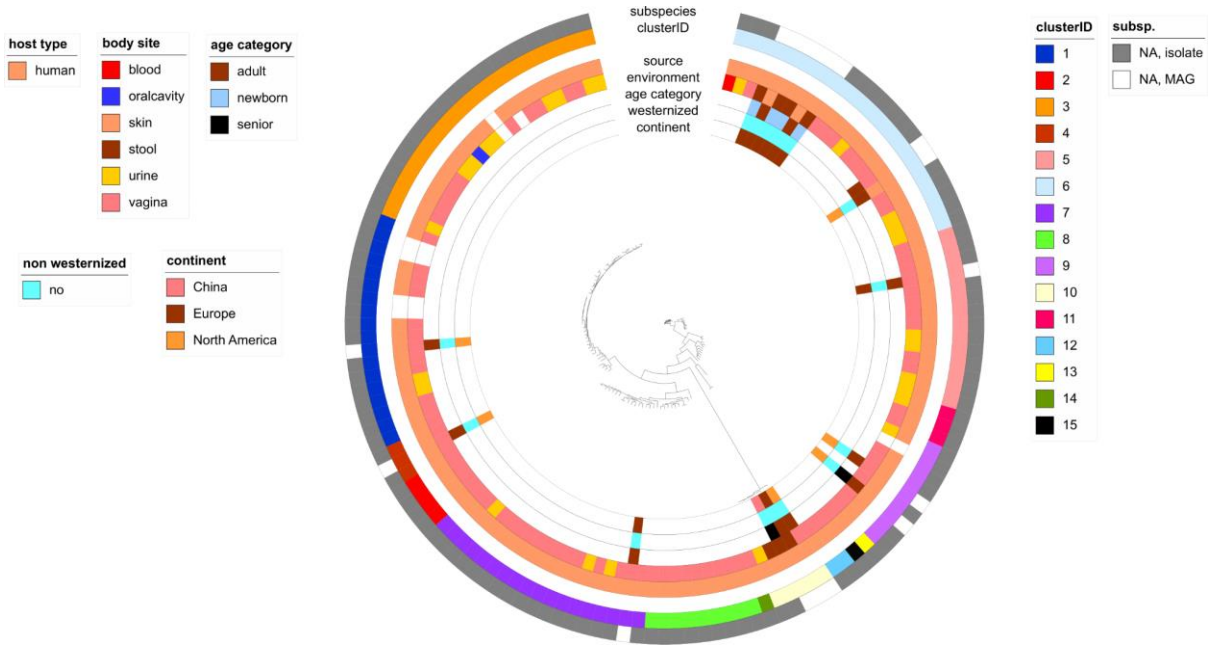
Supplementary Figure 4.6. Phylogenetic analysis for *B. dentium* by integrating isolate genomes and MAGs. The phylogenetic tree reports multiple host characteristics.



Supplementary Figure 4.7. Phylogenetic analysis for *B. pseudocatenulatum* by integrating isolate genomes and MAGs. The phylogenetic tree reports multiple host characteristics.



Supplementary Figure 4.8. Phylogenetic analysis for *B. pseudolongum* by integrating isolate genomes and MAGs. The phylogenetic tree reports multiple host characteristics.



Supplementary Figure 4.9. Phylogenetic analysis for *G. vaginalis* by integrating isolate genomes and MAGs.

The phylogenetic tree reports multiple host characteristics. *G. vaginalis* spans 10 SGBs that are further divided into different subspecies for a total of 15 clusters.

Supplementary Table 3.1. Description of the food metagenomes collected in this study

Supplementary Table 3.2. Taxonomic profiles generated with MetaPhlAn3 with curated metadata of all metagenomic samples from human microbiomes considered in this paper

Supplementary Table 3.3. Average prevalence and relative abundance of the 30 LAB species across the 303 food metagenomes estimated with MetaPhlAn3.

Supplementary Table 3.4. P-values associated with the Fisher's test (after FDR correction) applied on the presence/absence of the 30 LAB species determined by MetaPhlAn3 spanning age, body site, continent, and westernized lifestyle categories.

Supplementary Table 3.5. Description of the MAGs extracted in this study from food metagenomes with their assigned SGB and estimated taxonomy.

Supplementary Table 3.6. Summary of the SGBs retrieved from food metagenomes in this study, and their overlap with SGBs extracted from human metagenomes.

Supplementary Table 3.7. Source type for the reference genomes associated with the LAB species considered in this study.

Supplementary Table 3.8. Description of the MAGs extracted from non-human primate metagenomes with their assigned SGB and estimated taxonomy. pSGBs (primate SGBs) identify newly assembled SGBs from NHPs metagenomes only.

Supplementary Table 3.9. List of sugar metabolism genes found to be differently prevalent ($p < 0.05$) between food and human gut genomes in S. thermophilus and Lc. lactis.

Supplementary Table 3.10. Number of markers per species used to build the phylogenies through PhyloPhlAn.

Supplementary Table 4.1. List and description of the 54 publicly available datasets associated with human metagenomes. For each dataset, we report the number of samples and the PMID of the original publication in addition to relevant metadata information (i.e., age category, environment, lifestyle, source, and study condition).

Supplementary Table 4.2. Summary of the 1,192 genomes from isolate sources considered in this study. For each genome, we report basic statistics (i.e., genome length and GC content) in addition to metadata information in terms of environment retrieved from the NCBI portal or from original publications. For the human genomes, we also report the age category and source of the host. Finally, we report the taxonomy as reported in the NCBI portal in addition to the SGBID assigned by our computational pipeline aiming at delineating SGBs and clustering genomes into them.

Supplementary Table 4.3. Relative abundance for taxa belonging to the Bifidobacteriaceae family along with metadata information for the 9,528 human metagenomes considered in this study.

Supplementary Table 4.4. Average relative abundance and prevalence for taxa belonging to the Bifidobacteriaceae family for the 9,528 human metagenomes considered in this study. Values are stratified by age category, bodysite, and lifestyle.

Supplementary Table 4.5. Summary of the 110 SGBs belonging to Bifidobacteriaceae identified from the set of isolate genomes and MAGs considered in this study. For each SGB, we report the number of genomes retrieved from both isolate and metagenomic sources.

7 Bibliography

1. Berg G, Rybakova D, Fischer D, Cernava T, Vergès M-CC, Charles T, et al. Microbiome definition re-visited: old concepts and new challenges. Microbiome. 2020. doi:10.1186/s40168-020-00875-0

2. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007;449: 804–810.

3. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nat Med. 2018;24: 392–400.

4. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet. 2012;13: 260–270.

5. Ursell LK, Metcalf JL, Parfrey LW, Knight R. Defining the human microbiome. Nutr Rev. 2012;70 Suppl 1: S38–44.

6. Aagaard K, Luna RA, Versalovic J. The Human Microbiome of Local Body Sites and Their Unique Biology. Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases. 2015. pp. 11–18. doi:10.1016/b978-1-4557-4801-3.00002-3

7. Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJM, et al. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci U S A. 2015;112: E2930–8.

8. De Filippis F, Parente E, Ercolini D. Recent Past, Present, and Future of the Food Microbiome. Annu Rev Food Sci Technol. 2018;9: 589–608.

9. Sabater C, Cobo-Díaz JF, Álvarez-Ordóñez A, Ruas-Madiedo P, Ruiz L, Margolles A. Novel methods of microbiome analysis in the food industry. Int Microbiol. 2021;24: 593–605.

10. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35: 833–844.

11. McHugh AJ, Feehily C, Fenelon MA, Gleeson D, Hill C, Cotter PD. Tracking the Dairy Microbiota from Farm Bulk Tank to Skimmed Milk Powder. mSystems. 2020;5. doi:10.1128/mSystems.00226-20

12. Bovo S, Utzeri VJ, Ribani A, Cabbri R, Fontanesi L. Shotgun sequencing of honey DNA can describe honey bee derived environmental signatures and the honey bee hologenome complexity. Sci Rep. 2020;10: 9279.

13. Suárez N, Weckx S, Minahk C, Hebert EM, Saavedra L. Metagenomics-based approach for studying and selecting bioprotective strains from the bacterial community of artisanal cheeses. Int J Food Microbiol. 2020;335: 108894.

14. Walsh AM, Macori G, Kilcawley KN, Cotter PD. Meta-analysis of cheese microbiomes

highlights contributions to multiple aspects of quality. Nature Food. 2020;1: 500–510.

15. Ferrocino I, Bellio A, Giordano M, Macori G, Romano A, Rantsiou K, et al. Shotgun Metagenomics and Volatilome Profile of the Microbiota of Fermented Sausages. Appl Environ Microbiol. 2018;84. doi:10.1128/AEM.02120-17

16. Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008;9: 387–402.

17. Franzosa EA, Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, et al. Sequencing and beyond: integrating molecular "omics" for microbial community profiling. Nature Reviews Microbiology. 2015. pp. 360–372. doi:10.1038/nrmicro3451

18. Norman JM, Handley SA, Virgin HW. Kingdom-agnostic metagenomics and the importance of complete characterization of enteric microbial communities. Gastroenterology. 2014;146: 1459–1469.

19. Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell. 2015;160: 447–460.

20. Yang C, Chowdhury D, Zhang Z, Cheung WK, Lu A, Bian Z, et al. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. Comput Struct Biotechnol J. 2021;19: 6301–6314.

21. Lin H-H, Liao Y-C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. Scientific Reports. 2016. doi:10.1038/srep24175

22. Wang Z, Wang Z, Lu YY, Sun F, Zhu S. SolidBin: improving metagenome binning with semi-supervised normalized cut. Bioinformatics. 2019;35: 4229–4238.

23. Yu G, Jiang Y, Wang J, Zhang H, Luo H. BMC3C: binning metagenomic contigs using codon usage, sequence composition and read coverage. Bioinformatics. 2018;34: 4172–4179.

24. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell. 2019;176: 649–662.e20.

25. Van Rossum T, Ferretti P, Maistrenko OM, Bork P. Diversity within species: interpreting strains in microbiomes. Nat Rev Microbiol. 2020;18: 491–506.

26. Brown TA. Mapping Genomes. Wiley-Liss; 2002.

27. Lawrence JG, Retchless AC. The interplay of homologous recombination and horizontal gene transfer in bacterial speciation. Methods Mol Biol. 2009;532: 29–53.

28. Sheppard SK, Guttman DS, Fitzgerald JR. Population genomics of bacterial host adaptation. Nat Rev Genet. 2018;19: 549–565.

29. Bobay L-M, Ochman H. Factors driving effective population size and pan-genome evolution in bacteria. BMC Evol Biol. 2018;18: 153.

30. Shapiro BJ. Population genomics: microorganisms. Springer; 2018.

31. Raphael BJ. Research in Computational Molecular Biology: 22nd Annual International Conference, RECOMB 2018, Paris, France, April 21-24, 2018, Proceedings. Springer; 2018.

32. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9: 5114.

33. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17: 132.

34. Tibshirani R, Walther G. Cluster Validation by Prediction Strength. J Comput Graph Stat. 2005;14: 511–528.

35. Costea PI, Coelho LP, Sunagawa S, Munch R, Huerta-Cepas J, Forslund K, et al. Subspecies in the global human gut microbiome. Mol Syst Biol. 2017;13: 960.

36. Costea PI, Munch R, Coelho LP, Paoli L, Sunagawa S, Bork P. metaSNV: A tool for metagenomic strain level analysis. PLoS One. 2017;12: e0182392.

37. Mende DR, Sunagawa S, Zeller G, Bork P. Accurate and universal delineation of prokaryotic species. Nat Methods. 2013;10: 881–884.

38. Tett A, Huang KD, Asnicar F, Fehlner-Peach H, Pasolli E, Karcher N, et al. The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. Cell Host Microbe. 2019;26: 666–679.e7.

39. Karcher N, Pasolli E, Asnicar F, Huang KD, Tett A, Manara S, et al. Analysis of 1321 Eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. Genome Biol. 2020;21: 138.

40. Hennig C. Cluster-wise assessment of cluster stability. Comput Stat Data Anal. 2007;52: 258–271.

41. Gower JC. Principal Coordinates Analysis. Wiley StatsRef: Statistics Reference Online. Wiley; 2014. doi:10.1002/9781118445112.stat05670

42. Hennig C. Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods. J Multivar Anal. 2008;99: 1154–1176.

43. Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons; 2009.

44. Cordain L, Eaton SB, Sebastian A, Mann N, Lindeberg S, Watkins BA, et al. Origins and evolution of the Western diet: health implications for the 21st century. Am J Clin Nutr. 2005;81: 341–354.

45. Marco ML, Heeney D, Binda S, Cifelli CJ, Cotter PD, Foligné B, et al. Health benefits of fermented foods: microbiota and beyond. Curr Opin Biotechnol. 2017;44: 94–102.

46. Stefanovic E, Fitzgerald G, McAuliffe O. Advances in the genomics and metabolomics of dairy lactobacilli: A review. Food Microbiol. 2017;61: 33–49.

47. Wu C, Huang J, Zhou R. Genomics of lactic acid bacteria: Current status and potential applications. Crit Rev Microbiol. 2017;43: 393–404.

48. Douillard FP, de Vos WM. Biotechnology of health-promoting bacteria. Biotechnol Adv. 2019;37: 107369.

49. Hill C, Guarner F, Reid G, Gibson GR, Merenstein DJ, Pot B, et al. Expert consensus document. The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic. Nat Rev Gastroenterol Hepatol. 2014;11: 506–514.

50. Derrien M, van Hylckama Vlieg JET. Fate, activity, and impact of ingested bacteria within the human gut microbiota. Trends Microbiol. 2015;23: 354–366.

51. Bertuzzi AS, Walsh AM, Sheehan JJ, Cotter PD, Crispie F, McSweeney PLH, et al. Omics-Based Insights into Flavor Development and Microbial Succession within Surface-Ripened Cheese. mSystems. 2018;3. doi:10.1128/mSystems.00211-17

52. Escobar-Zepeda A, Sanchez-Flores A, Quirasco Baruch M. Metagenomic analysis of a Mexican ripened cheese reveals a unique complex microbiota. Food Microbiol. 2016;57: 116–127.

53. Milani C, Duranti S, Napoli S, Alessandri G, Mancabelli L, Anzalone R, et al. Colonization of the human gut by bovine bacteria present in Parmesan cheese. Nat Commun. 2019;10: 1286.

54. Quigley L, O'Sullivan DJ, Daly D, O'Sullivan O, Burdikova Z, Vana R, et al. Thermus and the Pink Discoloration Defect in Cheese. mSystems. 2016;1. doi:10.1128/mSystems.00023-16

55. Walsh AM, Crispie F, Kilcawley K, O'Sullivan O, O'Sullivan MG, Claesson MJ, et al. Microbial Succession and Flavor Production in the Fermented Dairy Beverage Kefir. mSystems. 2016;1. doi:10.1128/mSystems.00052-16

56. Walsh AM, Crispie F, Daari K, O'Sullivan O, Martin JC, Arthur CT, et al. Strain-Level Metagenomic Analysis of the Fermented Dairy Beverage Nunu Highlights Potential Food Safety Risks. Appl Environ Microbiol. 2017;83. doi:10.1128/AEM.01144-17

57. Wolfe BE, Button JE, Santarelli M, Dutton RJ. Cheese rind communities provide tractable systems for in situ and in vitro studies of microbial diversity. Cell. 2014;158: 422–433.

58. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, et al. Accessible, curated metagenomic data through ExperimentHub. Nat Methods. 2017;14: 1023–1024.

59. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife. 2021;10. doi:10.7554/eLife.65088

60. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27: 824–834.

61. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28: 1420–1428.

62. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7: e7359.

63. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25: 1043–1055.

64. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017;35: 725.

65. Asnicar F, Thomas AM, Beghini F. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn2. Nat Commun. under review.

66. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410.

67. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30: 772–780.

68. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25: 1972–1973.

69. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5: e9490.

70. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30: 1312–1313.

71. Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ. 2015. p. e1029. doi:10.7717/peerj.1029

72. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30: 2068–2069.

73. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31: 3691–3693.

74. Redondo-Useros N, Gheorghe A, Díaz-Prieto LE, Villavisencio B, Marcos A, Nova E. Associations of Probiotic Fermented Milk (PFM) and Yogurt Consumption with Bifidobacterium and Lactobacillus Components of the Gut Microbiota in Healthy Adults. Nutrients. 2019;11. doi:10.3390/nu11030651

75. Vatanen T, Kostic AD, d'Hennezel E, Siljander H, Franzosa EA, Yassour M, et al. Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. Cell. 2016;165: 1551.

76. Soto A, Martín V, Jiménez E, Mader I, Rodríguez JM, Fernández L. Lactobacilli and bifidobacteria in human breast milk: influence of antibiotherapy and other host and clinical factors. J Pediatr Gastroenterol Nutr. 2014;59: 78–88.

77. Murphy K, Curley D, O'Callaghan TF, O'Shea C-A, Dempsey EM, O'Toole PW, et al. The Composition of Human Milk and Infant Faecal Microbiota Over the First Three Months of Life: A Pilot Study. Sci Rep. 2017;7: 40597.

78. Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, Fedi S, et al. Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host Microbe. 2018;24: 133–145.e5.

79. Di Cagno R, Coda R, De Angelis M, Gobbetti M. Exploitation of vegetables and fruits through lactic acid fermentation. Food Microbiol. 2013;33: 1–10.

80. Broussard JL, Devkota S. The changing microbial landscape of Western society: Diet, dwellings and discordance. Mol Metab. 2016;5: 737–742.

81. Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, et al. Gut microbiome of the Hadza hunter-gatherers. Nat Commun. 2014;5: 3654.

82. Crittenden AN, Schnorr SL. Current views on hunter-gatherer nutrition and the evolution of the human diet. Am J Phys Anthropol. 2017;162: 84–109.

83. Zhu K, Tan F, Mu J, Yi R, Zhou X, Zhao X. Anti-Obesity Effects of Lactobacillus fermentum CQPC05 Isolated from Sichuan Pickle in High-Fat Diet-Induced Obese Mice through PPAR-α Signaling Pathway. Microorganisms. 2019;7. doi:10.3390/microorganisms7070194

84. Manara S, Asnicar F, Beghini F, Bazzani D, Cumbo F, Zolfo M, et al. Microbial genomes from non-human primate gut metagenomes expand the primate-associated bacterial tree of life with over 1000 novel species. Genome Biol. 2019;20: 299.

85. Roos S, Karner F, Axelsson L, Jonsson H. Lactobacillus mucosae sp. nov., a new species with in vitro mucus-binding activity isolated from pig intestine. Int J Syst Evol Microbiol. 2000;50 Pt 1: 251–258.

86. Heredia N, García S. Animals as sources of food-borne pathogens: A review. Anim Nutr. 2018;4: 250–255.

87. Martin NH, Trmčić A, Hsieh T-H, Boor KJ, Wiedmann M. The Evolving Role of Coliforms As Indicators of Unhygienic Processing Conditions in Dairy Foods. Frontiers in Microbiology. 2016. doi:10.3389/fmicb.2016.01549

88. Zanine A de M, Bonelli EA, de Souza AL, Ferreira D de J, Santos EM, Ribeiro MD, et al. Effects of Streptococcus bovis Isolated from Bovine Rumen on the Fermentation Characteristics and Nutritive Value of Tanzania Grass Silage. ScientificWorldJournal. 2016;2016: 8517698.

89. Beck M, Frodl R, Funke G. Comprehensive Study of Strains Previously Designated Streptococcus bovis Consecutively Isolated from Human Blood Cultures and Emended Description of Streptococcus gallolyticus and Streptococcus infantarius subsp. coli. Journal of Clinical Microbiology. 2008. pp. 2966–2972. doi:10.1128/jcm.00078-08

90. Godon JJ, Delorme C, Ehrlich SD, Renault P. Divergence of Genomic Sequences between Lactococcus lactis subsp. lactis and Lactococcus lactis subsp. cremoris. Appl Environ Microbiol. 1992;58: 4045–4047.

91. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new genomic blueprint of the human gut microbiota. Nature. 2019;568: 499–504.

92. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. New insights from uncultivated genomes of the global human gut microbiome. Nature. 2019;568: 505–510.

93. Segers ME, Lebeer S. Towards a better understanding of Lactobacillus rhamnosus GG--host interactions. Microb Cell Fact. 2014;13 Suppl 1: S7.

94. Wuyts S, Wittouck S, De Boeck I, Allonsius CN, Pasolli E, Segata N, et al. Large-Scale Phylogenomics of the Lactobacillus casei Group Highlights Taxonomic Inconsistencies and Reveals Novel Clade-Associated Features. mSystems. 2017;2. doi:10.1128/mSystems.00061-17

95. Li X, Liang S, Xia Z, Qu J, Liu H, Liu C, et al. Establishment of a Macaca fascicularis gut microbiome gene catalog and comparison with the human, pig, and mouse gut microbiomes. Gigascience. 2018;7. doi:10.1093/gigascience/giy100

96. Zheng J, Wittouck S, Salvetti E, Franz CMAP, Harris HMB, Mattarelli P, et al. A taxonomic note on the genus Lactobacillus: Description of 23 novel genera, emended description of the genus Lactobacillus Beijerinck 1901, and union of Lactobacillaceae and Leuconostocaceae. Int J Syst Evol Microbiol. 2020;70: 2782–2858.

97. Uriot O, Denis S, Junjua M, Roussel Y, Dary-Mourot A, Blanquet-Diot S. Streptococcus thermophilus: From yogurt starter to a new promising probiotic candidate? J Funct Foods. 2017;37: 74–89.

98. Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen T, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science. 2016;352: 565–569.

99. Song AA-L, In LLA, Lim SHE, Rahim RA. A review on Lactococcus lactis: from food to factory. Microb Cell Fact. 2017;16: 55.

100. Nakamura S, Morimoto YV, Kudo S. A lactose fermentation product produced by Lactococcus lactis subsp. lactis, acetate, inhibits the motility of flagellated pathogenic bacteria. Microbiology. 2015;161: 701–707.

101. Food and Agriculture Organization of the United Nations. New Food Balances. Available: http://www.fao.org/faostat/en/#data/FBS

102. Taylor BC, Lejzerowicz F, Poirel M, Shaffer JP, Jiang L, Aksenov A, et al. Consumption of Fermented Foods Is Associated with Systematic Differences in the Gut Microbiome and Metabolome. mSystems. 2020;5. doi:10.1128/mSystems.00901-19

103. Jung JY, Lee SH, Kim JM, Park MS, Bae J-W, Hahn Y, et al. Metagenomic analysis of kimchi, a traditional Korean fermented food. Appl Environ Microbiol. 2011;77: 2264–2274.

104. Tamang JP, Watanabe K, Holzapfel WH. Review: Diversity of microorganisms in global fermented foods and beverages. Front Microbiol. 2016; 7: 377. 2016.

105. Prentice AM. Dairy products in global public health. Am J Clin Nutr. 2014;99: 1212S–6S.

106. Thomas AM, Segata N. Multiple levels of the unknown in microbiome research. BMC Biol. 2019;17: 48.

107. Hemarajata P, Versalovic J. Effects of probiotics on gut microbiota: mechanisms of intestinal immunomodulation and neuromodulation. Therap Adv Gastroenterol. 2013;6: 39–51.

108. Mitsuoka T. Bifidobacteria and their role in human health. J Ind Microbiol. 1990;6: 263–267.

109. Turroni F, van Sinderen D, Ventura M. Genomics and ecological overview of the genus Bifidobacterium. Int J Food Microbiol. 2011;149: 37–44.

110. Arboleya S, Watkins C, Stanton C, Ross RP. Gut Bifidobacteria Populations in Human Health and Aging. Front Microbiol. 2016;7: 1204.

111. Turroni F, Ventura M, Buttó LF, Duranti S, O'Toole PW, Motherway MO, et al. Molecular dialogue between the human gut microbiota and the host: a Lactobacillus and Bifidobacterium perspective. Cell Mol Life Sci. 2014;71: 183–203.

112. Favier CF, Vaughan EE, De Vos WM, Akkermans ADL. Molecular monitoring of succession of bacterial communities in human neonates. Appl Environ Microbiol. 2002;68: 219–226.

113. Odamaki T, Kato K, Sugahara H, Hashikura N, Takahashi S, Xiao J-Z, et al. Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. BMC Microbiol. 2016;16: 90.

114. Scott KP, Jean-Michel A, Midtvedt T. Manipulating the gut microbiota to maintain health and treat disease. Microb Ecol Health Dis. 2015. Available: https://www.tandfonline.com/doi/abs/10.3402/mehd.v26.25877

115. Wong CB, Odamaki T, Xiao J-Z. Beneficial effects of Bifidobacterium longum subsp. longum BB536 on human health: Modulation of gut microbiome as the principal action. J Funct Foods. 2019;54: 506–519.

116. Ménard O, Butel M-J, Gaboriau-Routhiau V, Waligora-Dupriet A-J. Gnotobiotic mouse immune response induced by Bifidobacterium sp. strains isolated from infants. Appl Environ Microbiol. 2008;74: 660–666.

117. Di Gioia D, Aloisio I, Mazzola G, Biavati B. Bifidobacteria: their impact on gut microbiota composition and their applications as probiotics in infants. Appl Microbiol Biotechnol. 2014;98: 563–577.

118. Biagi E, Candela M, Fairweather-Tait S, Franceschi C, Brigidi P. Ageing of the human metaorganism: the microbial counterpart. Age. 2012;34: 247–267.

119. Malaguarnera G, Leggio F, Vacante M, Motta M, Giordano M, Bondi A, et al. Probiotics in the gastrointestinal diseases of the elderly. J Nutr Health Aging. 2012;16: 402–410.

120. Tojo R, Suárez A, Clemente MG, de los Reyes-Gavilán CG, Margolles A, Gueimonde M, et al. Intestinal microbiota in health and disease: role of bifidobacteria in gut homeostasis. World J Gastroenterol. 2014;20: 15163–15176.

121. Guardamagna O, Amaretti A, Puddu PE, Raimondi S. Bifidobacteria supplementation: effects on plasma lipid profiles in dyslipidemic children. Nutrition. 2014. Available: https://www.sciencedirect.com/science/article/pii/S089990071400080X?casa_token=aifAIwpiJ6EAAAAA:VR7abDlXxgnmSZozRt-agdUmioOP3xVZZQXaGikM96QWoQATcsdOiJlvdU36fFSRIk8k4KN6juI

122. Bercik P, Park AJ, Sinclair D, Khoshdel A, Lu J, Huang X, et al. The anxiolytic effect of Bifidobacterium longum NCC3001 involves vagal pathways for gut-brain communication. Neurogastroenterol Motil. 2011;23: 1132–1139.

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

124. Asnicar F, Thomas AM, Beghini F, Mengoni C, Manara S, Manghi P, et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nat Commun. 2020;11: 2500.

125. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47: W256–W259.

126. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLoS Comput Biol. 2016;12: e1004977.

127. Castro-Nallar E, Bendall ML, Pérez-Losada M, Sabuncyan S, Severance EG, Dickerson FB, et al. Composition, taxonomy and functional diversity of the oropharynx microbiome in individuals with schizophrenia and controls. PeerJ. 2015;3: e1140.

128. Chng KR, Tay ASL, Li C, Ng AHQ, Wang J, Suri BK, et al. Whole metagenome profiling reveals skin microbiome-dependent susceptibility to atopic dermatitis flare. Nat Microbiol. 2016;1: 16106.

129. David LA, Weil A, Ryan ET, Calderwood SB, Harris JB, Chowdhury F, et al. Gut microbial succession follows acute secretory diarrhea in humans. MBio. 2015;6: e00381–15.

130. Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. Nat Commun. 2015;6: 1–13.

131. Ghensi P, Manghi P, Zolfo M, Armanini F, Pasolli E, Bolzan M, et al. Strong oral plaque microbiome signatures for dental implant diseases identified by strain-resolution metagenomics. NPJ Biofilms Microbiomes. 2020;6: 47.

132. Hannigan GD, Duhaime MB, Ruffin MT 4th, Koumpouras CC, Schloss PD. Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. MBio. 2018;9.

doi:10.1128/mBio.02248-18

133. Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat Microbiol. 2016;2: 16180.

134. Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature. 2013;498: 99–103.

135. Kieser S, Sarker SA, Sakwinska O, Foata F, Sultana S, Khan Z, et al. Bangladeshi children with acute diarrhoea show faecal microbiomes with increased*Streptococcus*abundance, irrespective of diarrhoea aetiology. Environmental Microbiology. 2018. pp. 2256–2269. doi:10.1111/1462-2920.14274

136. Kostic AD, Gevers D, Siljander H, Vatanen T, Hyötyläinen T, Hämäläinen A-M, et al. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. Cell Host Microbe. 2015;17: 260–273.

137. Li J, Zhao F, Wang Y, Chen J, Tao J, Tian G, et al. Gut microbiota dysbiosis contributes to the development of hypertension. Microbiome. 2017. doi:10.1186/s40168-016-0222-x

138. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014;32: 822–828.

139. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012;490: 55–60.

140. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature. 2014;513: 59–64.

141. Raymond F, Ouameur AA, Déraspe M, Iqbal N, Gingras H, Dridi B, et al. The initial state of the human gut microbiome determines its reshaping by antibiotics. ISME J. 2016;10: 707–720.

142. Vincent C, Miller MA, Edens TJ, Mehrotra S, Dewar K, Manges AR. Bloom and bust: intestinal microbiota dynamics in response to hospital exposures and Clostridium difficile colonization or infection. Microbiome. 2016. doi:10.1186/s40168-016-0156-3

143. Vogtmann E, Hua X, Zeller G, Sunagawa S, Voigt AY, Hercog R, et al. Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. PLoS One. 2016;11: e0155362.

144. Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut. 2017;66: 70–78.

145. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol. 2014;10: 766.