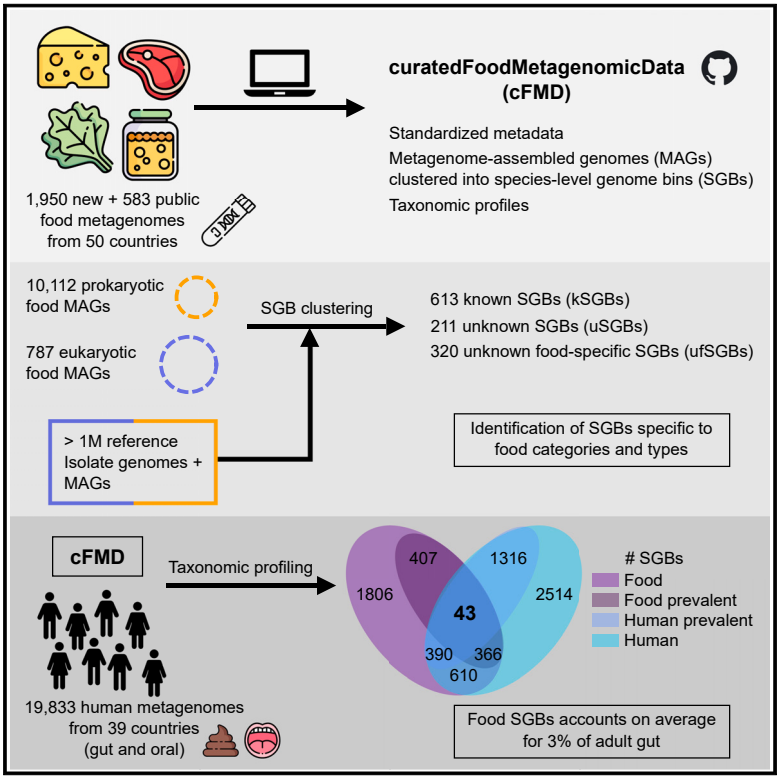
Cell

Unexplored microbial diversity from 2,500 food metagenomes and links with the human microbiome

Graphical abstract



Highlights

- With curatedFoodMetagenomicData, we integrated and analyzed >2,500 food metagenomes
- Over 10,000 prokaryotic and eukaryotic MAGs uncover substantial food microbial diversity
- Food microbes account for up to an average of 3% of the adult gut microbiome
- Strain-level analysis highlights potential instances of food-to-gut microbe transmission

Authors

Niccolò Carlino, Aitor Blanco-Míguez, Michal Punčochář, ..., Paul D. Cotter, Nicola Segata, Edoardo Pasolli

Correspondence

nicola.segata@unitn.it

In brief

Despite extensive studies on nutrition and the abiotic components of food, the food microbiome remains largely uncharacterized. In this study, a systematic metagenomic sequencing of microbiomes from over 2,500 food sources, integrated with metadata and analyzed in connection with the human microbiome, showed that many members of the gut microbiome may have been recently acquired from food sources.

Carlino et al., 2024, Cell *187*, 5775–5795
October 3, 2024
https://doi.org/10.1016/j.cell.2024.07.039

CellPress

Cell

CellPress
OPEN ACCESS

Resource

Unexplored microbial diversity from 2,500 food metagenomes and links with the human microbiome

Niccolò Carlino,[1] Aitor Blanco-Míguez,[1] Michal Punčochář,[1] Claudia Mengoni,[1] Federica Pinto,[1] Alessia Tatti,[2,3,4] Paolo Manghi,[1] Federica Armanini,[1] Michele Avagliano,[5] Coral Barcenilla,[6] Samuel Breselge,[7,8] Raul Cabrera-Rubio,[7,9] Inés Calvete-Torre,[10,11] Mairéad Coakley,[7] José F. Cobo-Díaz,[6] Francesca De Filippis,[5,12] Hrituraj Dey,[1] John Leech,[7] Eline S. Klaassens,[13] Stephen Knobloch,[14] Dominic O'Neil,[15] Narciso M. Quijada,[16,17,18] Carlos Sabater,[10,11] Sigurlaug Skírnisdóttir,[14] Vincenzo Valentino,[5] Liam Walsh,[7,8,19] MASTER EU Consortium, Avelino Alvarez-Ordóñez,[6] Francesco Asnicar,[1] Gloria Fackelmann,[1] Vitor Heidrich,[1] Abelardo Margolles,[10,11] Viggó Thór Marteinsson,[14,20] Omar Rota Stabelli,[3,4] Martin Wagner,[16,17] Danilo Ercolini,[5,12,24] Paul D. Cotter,[7,8,21,24] Nicola Segata,[1,22,23,24,25,*] and Edoardo Pasolli[5,12,24]

[1]Department of Cellular, Computational and Integrative Biology, University of Trento, Trento, Italy
[2]Scuola Universitaria Superiore IUSS Pavia, Pavia, Italy
[3]Centre for Agriculture Food Environment, University of Trento, Trento, Italy
[4]Research and Innovation Centre, Fondazione Edmund Mach, San Michele All'Adige, Italy
[5]Department of Agricultural Sciences, Division of Microbiology, University of Naples Federico II, Portici, Italy
[6]Department of Food Hygiene and Technology, Universidad de León, León, Spain
[7]Teagasc Food Research Centre, Moorepark, Fermoy, Co. Cork, Ireland
[8]APC Microbiome Ireland, University College Cork, Cork, Ireland
[9]Department of Biotechnology, Institute of Agrochemistry and Food Technology - National Research Council (IATA-CSIC), Paterna, Valencia, Spain
[10]Department of Microbiology and Biochemistry of Dairy Products, Instituto de Productos Lácteos de Asturias - Consejo Superior de Investigaciones Científicas (IPLA-CSIC), Villaviciosa, Spain
[11]Microhealth Group, Instituto de Investigación Sanitaria del Principado de Asturias (ISPA), Oviedo, Spain
[12]Task Force on Microbiome Studies, University of Naples Federico II, Portici, Italy
[13]BaseClear B.V., Leiden, the Netherlands
[14]Microbiology Research Group, Matís, Reykjavík, Iceland
[15]QIAGEN GmbH, Hilden, Germany
[16]Austrian Competence Centre for Feed and Food Quality, Safety, and Innovation, FFoQSI GmbH, Tulln an der Donau, Austria
[17]Unit of Food Microbiology, Institute of Food Safety, Food Technology and Veterinary Public Health, Department for Farm Animals and Veterinary Public Health, University of Veterinary Medicine Vienna, Vienna, Austria
[18]Institute for Agribiotechnology Research (CIALE), Department of Microbiology and Genetics, University of Salamanca, Salamanca, Spain
[19]School of Microbiology, University College Cork, Cork, Ireland
[20]University of Iceland, Faculty of Food Science and Nutrition, Reykjavík, Iceland
[21]VistaMilk SFI Research Centre, Teagasc, Moorepark, Fermoy, Co. Cork, Ireland
[22]IEO, Istituto Europeo di Oncologia IRCSS, Milan, Italy
[23]Department of Twins Research and Genetic Epidemiology, King's College London, London, UK
[24]Senior author
[25]Lead contact
*Correspondence: nicola.segata@unitn.it
https://doi.org/10.1016/j.cell.2024.07.039

SUMMARY

Complex microbiomes are part of the food we eat and influence our own microbiome, but their diversity remains largely unexplored. Here, we generated the open access curatedFoodMetagenomicData (cFMD) resource by integrating 1,950 newly sequenced and 583 public food metagenomes. We produced 10,899 metagenome-assembled genomes spanning 1,036 prokaryotic and 108 eukaryotic species-level genome bins (SGBs), including 320 previously undescribed taxa. Food SGBs displayed significant microbial diversity within and between food categories. Extension to >20,000 human metagenomes revealed that food SGBs accounted on average for 3% of the adult gut microbiome. Strain-level analysis highlighted potential instances of food-to-gut transmission and intestinal colonization (e.g., *Lacticaseibacillus paracasei*) as well as SGBs with divergent genomic structures in food and humans (e.g., *Streptococcus gallolyticus* and *Limosilactobabillus mucosae*). The cFMD expands our knowledge on food microbiomes, their role in shaping the human microbiome, and supports future uses of metagenomics for food quality, safety, and authentication.

INTRODUCTION

Microorganisms have had a fundamental role in the history of food science.[1] Humanity has always faced hazards due to microbial food poisoning and spoilage and preservation technologies (e.g., cooking, salting, and fermentation[2,3]) have been improved toward current standards of food safety, quality, and production yield.[4,5] Fermentation of raw plants, dairy, and meat is a dynamic process that can enhance the quality, diversity, and safety of food products by controlling potentially harmful bacteria and improving organoleptic properties[6] and health-promoting features.[7] Even in the absence of potentially pathogenic taxa, the level of biodiversity in foods is heterogeneous and ranges from single microbial associations (e.g., foods fermented with industrially selected starter cultures) to complex microbiomes.[8] The relevance of characterizing food-associated microbial communities to improve foods and understand their impact on human health is increasingly recognized, but much of this diversity remains unexplored.

In the pursuit of characterizing the microbial composition of food sources, the field received a boost when *in vitro* cultivation[9,10] was complemented by community-level molecular typing, initially using 16S rRNA gene sequencing[11] and currently also shotgun metagenomics.[12,13] While cultivation has extended to high-throughput settings,[14] shotgun metagenomics is the sole approach able to comprehensively survey the microbial composition of a sample and unearth its genomic potential[12,15] along with the reconstruction of genomes of multiple strains.[16,17] Nevertheless, metagenomic studies have been mainly conducted on single food types[18–24] and on limited sample sizes,[25,26] hindering inter-study integrative analyses currently available only for other environments such as the human gut.[27–31]

Studying the microbial content of food is crucial also in light of its impact on human health.[32–34] Microbiome-mediated links between diet and health have been prevalently investigated by considering the role of the abiotic components of foods.[35–41] However, the recently described horizontal and vertical person-to-person microbial transmission[42] is likely not the only source of microbiome diversity,[43] and early evidence of dietary microbes becoming human microbiome members motivates a more comprehensive investigation of this phenomenon across food types and human populations.[44–47] Comprehensive investigations of the overlap between food and human microbiomes at a population-scale and strain-level resolution are thus required.

Here, we present curatedFoodMetagenomicData (cFMD), an open-access resource that collects food-associated microbial data to support the use of metagenomics in food science. The current release comprises 2,533 food metagenomes with standardized metadata, 1,950 of them newly sequenced within the MASTER EU Consortium. We generated 10,112 prokaryotic and 787 eukaryotic metagenome-assembled genomes (MAGs) from food that were grouped into 1,036 prokaryotic and 108 eukaryotic species clusters, 320 of which resulted to be uncharacterized when compared with >1 M existing genomes. We included these MAGs into our pipelines for sensitive taxonomic profiling and applied it to 19,833 human metagenomes, revealing species- and strain-level overlaps along the food-human axis.

RESULTS

A compendium of 2,533 food-associated metagenomes for integrative microbiome analysis

To better investigate the microbial populations present in food, we acquired and metagenomically sequenced 1,950 food microbiomes (MASTER EU Consortium) and integrated them with 583 publicly available samples[16,18–20,22–26,45,46,48–80] (Figure 1A). This resulted in 2,533 metagenomes spanning 59 datasets (Table S1) from 50 countries (Figure 1B) collected over the last decade (Figure 1D). Samples from MASTER expanded the number of food metagenomes by 334% and significantly increased the sequencing depth (mean ± SD: 6.7 Gb ± 6.3 Gb/sample for MASTER and 3.0 Gb ± 5.8 Gb/sample for non-MASTER samples; Wilcoxon rank sum test $p < 0.001$). The multi-level metagenomic profiles obtained with advanced validated pipelines applied on all samples were collected in cFMD along with standardized metadata (Figure 1A; STAR Methods).

Metadata (Table S1) were organized into 27 fields, covering sample, food-related, and technical information (STAR Methods; syntactic rules defined in Table S1). Hierarchical food categorization of the samples was based on the food type/substrate, production approach (fermented/non-fermented), and other specific features of different food types[81,82] (Table S1), resulting in 15 top-level categories (Figure 1C), 107 types, and 358 subtypes. Although the majority of samples came from dairy sources (n = 1,650), fermented beverages (n = 422), and fermented meat (n = 133), we also considered less characterized categories such as fermented seeds, non-fermented fish, and non-fermented meat.

We performed standardized sample pre-processing, taxonomic and functional community profiling, and genomic reconstruction of single taxa (Figure 1A; STAR Methods). We generated 27,123 MAGs that, after quality filtering, resulted in 4,976 high-quality (HQ) and 5,136 medium-quality (MQ) prokaryotic MAGs[83] (Table S2). These MAGs were integrated with over 1 M genomic sequences[84] (including 173,302 isolate genomes, hereafter called "reference genomes")[28,84] and clustered at 95% whole-genome average nucleotide identity (ANI)[85] into 1,036 species-level genome bins (SGBs; hereafter "food SGBs" since comprising at least one MAG from food). We further identified SGB-specific marker genes for MetaPhlAn v4[84,86] (Figure 1E) to enable the profiling—even at low abundances—of all SGBs in metagenomes. We also reconstructed 392 HQ and 395 MQ eukaryotic MAGs that were clustered into 108 eukaryotic food SGBs. Such generated datasets were the basis of downstream analyses, which were extended to 19,833 human metagenomes available in curatedMetagenomicData (cMD)[27] to link food and human microbiomes (Figure 1A).

Broadening the phylogenetic diversity of food-associated bacterial species

The 10,112 prokaryotic MAGs clustered into 1,036 SGBs belonging to 13 different phyla, which we analyzed to assess the phylogenetic diversity of food microbes (Table S2; Figure 2A). Six classes from four phyla were primarily responsible for the expanded phylogenetic diversity as they comprised 92% of MAGs and 78% of SGBs coming from our study: Actinomycetia (Figure S1A), α-, β-, and ɣ-Proteobacteria (Figure S1B),

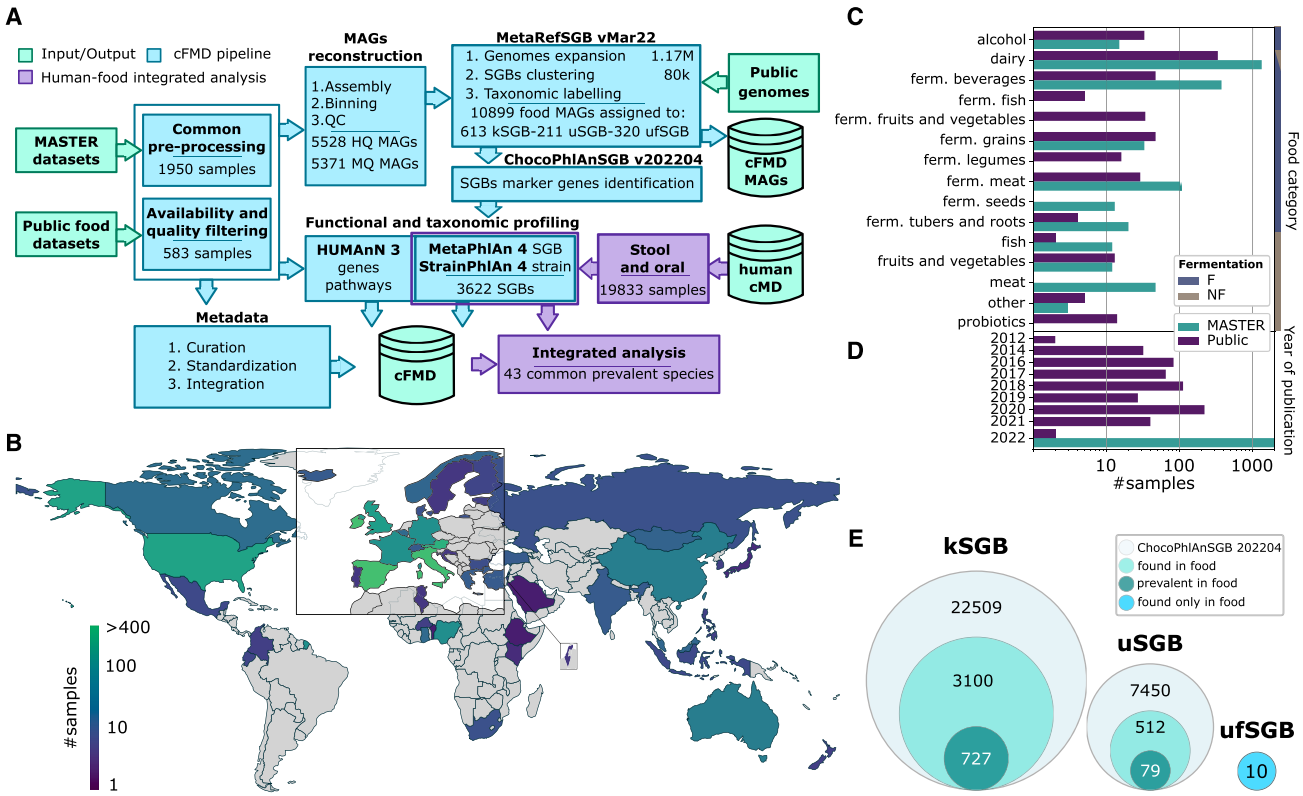


Figure 1. curatedFoodMetagenomicData (cFMD) provides >2,500 food metagenomes with standardized metadata and processed data

(A–D) (A) The main steps of the pipeline: (i) surveying and filtering of public data along with sequencing of samples within the MASTER EU Consortium from food sources; (ii) curation and standardization of metadata; (iii) generation of MAGs and taxonomic and functional profiles; (iv) integration of food with human metagenomes; and (v) release of public databases for downstream analyses. We enlarged the number of available samples (B) around the globe (i.e., from 51 countries), (C) across 15 food categories spanning fermented (F) and non-fermented (NF) foods (Table S1), and (D) in time. Such a resource was used to increase the set of MAGs and SGBs available in our databases.

(E) The number of SGBs available in the pipeline and detected by taxonomic profiling in at least one food metagenome, grouped by unknownness level (kSGB, known SGB; uSGB, unknown SGB; ufSGB, unknown food-specific SGB).

Flavobacteria (Figure S1C), and Bacilli (Figure S1D). The latter belongs to the phylum Bacillota (formerly Firmicutes) that comprised the majority of the food MAGs (n = 6,300, from 394 SGBs); this included lactic acid bacteria (LAB, in particular the order Lactobacillales[87]; 5,577 MAGs from 231 SGBs), which comprised the two most reconstructed families: *Lactobacillaceae* (3,447 MAGs) and *Streptococcaceae* (1,805 MAGs). 20% of the total food MAGs (n = 2,026 from 208 SGBs, Figure S1A) were associated with Actinomycetota (formerly Actinobacteria) and especially the class Actynomicetia (1,544 MAGs from 150 SGBs), including health-associated species belonging to the genera *Bifidobacterium* (168 MAGs from 10 SGBs) and *Propionibacterium* (37 MAGs from 2 SGBs). *Acetobacteraceae*, which encompass acetic acid bacteria (AAB), was the third most reconstructed family overall (523 MAGs from 61 SGBs) and represented a large portion of Pseudomonadota (formerly Proteobacteria; 1,652 MAGs from 366 SGBs; Figure S1B). Only 93 MAGs from 45 SGBs were assigned to the more typically gut-associated phylum Bacteroidota (formerly Bacteroidetes; Figure S1C), with class Flavobacteriia—occasionally associated with food spoilage[88]—being represented with 21 SGBs. Thus, cFMD provided an expanded genetic and microbial diversity of food microbes, and this included the relatively few bacterial families traditionally associated with food products that we next harnessed for in-depth investigations.

Expanding the genomic diversity of typical food-associated bacterial species

Half of the food SGBs (535 out of 1,036) contained at least one reference genome and were therefore assigned taxonomically at species level (known SGBs or known SGBs [kSGBs]; Figure 2B). 67% of the 7,961 MAGs from kSGBs were retrieved from dairy (5,334 MAGs from 312 kSGBs; Figures S2E and S2F), reflecting the 65% of samples derived from dairy products. The most reconstructed species from dairy (Figure 3A) were *Lactococcus lactis* (672 MAGs), *Streptococcus thermophilus* (448), *Lacticaseibacillus paracasei* (415), and *Lactococcus cremoris* (n = 404), and LABs represented 12 of the 15 largest SGBs overall. Non-LAB species prevalent in dairy were *Staphlyococcus equorum* (171 MAGs; Figure S1D), *Brevibacterium aurantiacum* (156; Figure S1A), *Corynebacterium casei* (89; Figure S1A), *Brevibacterium yomogidense* (68), and *Flaviflexus ciconiae* (68, isolated previously only from *Ciconia boyciana*[89]), many of which are usually found in

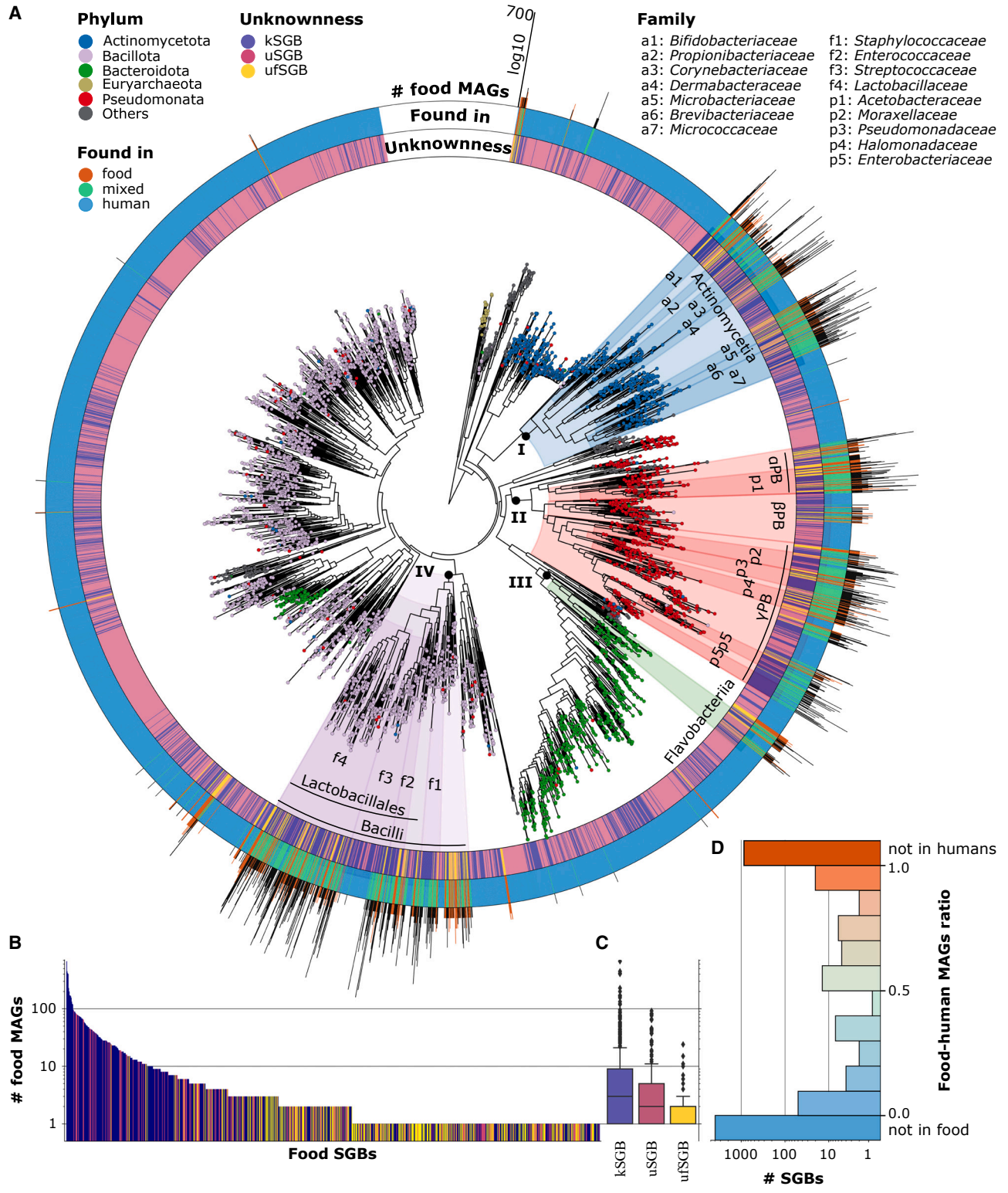


Figure 2. Phylogenetic tree of the 1,036 prokaryotic SGBs detected in food metagenomes

(A) Phylogeny with the SGBs reconstructed from food and integrated with 3,962 prokaryotic SGBs prevalent in human metagenomes. Each leaf represents an SGB and is colored according to phylum. Families prevalent in food are highlighted (a1-p5), and trees for relevant clades are reported in Figure S1: Actinomycetota (I); Pseudomonadota (II); Bacteroidota (III); and Bacillota (IV).

(legend continued on next page)

cheese-producing environments. Thus, the fact that the majority of samples were derived from dairy is reflected in their high representation among the most reconstructed kSGBs.

Non-dairy samples resulted in 2,651 MAGS and 331 kSGBs (62% of total kSGBs; Figure 3C). *L. paracasei* was the species with the highest number of non-dairy MAGs (238), which were retrieved across six categories and 27% of non-dairy samples overall. *Lactiplantibacillus plantarum* was the largest SGB in probiotic commercial products (50%; 7 MAGs), fermented grains (24%; 19), and fermented fruits and vegetables (18%; 6). In fermented foods, other prevalent SGBs included *Bacillus subtilis* in fermented seeds (77%), *Limosilactobacillus fermentum* in fermented tubers and roots (46%), *Weissella confusa* in fermented legumes (44%), *Latilactobacillus sakei* in fermented meat (31%), and *Tatumella ptyseos* in alcoholic beverages (19%). In general, non-fermented categories (i.e., fish, fruit and vegetables, meat, and other) did not yield any of the most frequently reconstructed SGBs (Figures 3C and 3D) but exhibited category-specific SGBs such as the spoilage-associated *Brochothrix thermosphacta* in meat (40%).[90] However, raw milk, considered a non-fermented food type in the dairy category (Table S1), was an exception since a few MAGs were retrieved for instance from *L. mesenteroides* (3 MAGs), *L. lactis* (3), *S. thermophilus* (2), and *L. cremoris* (1). Raw milk contains LAB that is abundant in fermented products; however, their MAG extraction is not trivial due to their lower abundance in the unfermented substrate.

Our database also expanded the genomic characterization of understudied species, considering, for example, that 18 of the 30 largest kSGBs comprised <10 reference genomes, such as *Lactobacillus kefiranofaciens* (74 MAGs), *Lactococcus laudensis* (59), *Lactococcus raffinolactis* (51), and *Lentilactobacillus otakiensis* (48). This applied also to SGBs prevalent in non-dairy such as *Liquorilactobacillus satsumensis* (200 MAGs; 6th largest SGB overall), *Liquorilactobacillus nagelii* (119 MAGs; 13th largest SGB) (Figure 3D), and *Acetobacter orientalis* (89 MAGs mostly from water and milk kefir, Figure S1B). Notably, *Zymomonas mobilis* (115 MAGs), one of the few known bacterial species capable of ethanol fermentation, was mainly retrieved from water kefir and was split across three kSGBs (i.e., SGB19526 corresponding to *Z. mobilis* in the Genome Taxonomy Database (GTDB),[91] SGB19527—*Z. pomaceae*, and SGB77042—*Z. mobilis_B*; Figures 3D and S1B).

Overall, for the bacterial species with isolated representatives found in food metagenomes, our integration with MAGs provides a much increased strain diversity that enables higher-resolution investigations.

Half of food microbes represent uncultured, unexplored species

We next investigated the other half of the food SGBs (501 out of 1,036) that did not contain any reference genomes and thus were considered unknown SGBs (uSGBs). Their overall diversity was very high with a 95% increase in the total branch length when added to the kSGB phylogeny (Figure 2A). These uSGBs contained 2,127 food MAGs (21% of the total; Figures 2B and 2C), and 46% of them were unassigned even at the genus level (<85% ANI to closest reference genomes[28]).

Such uncharacterized species were widespread as they were detected in all food categories (except in non-fermented fish having only 14 samples) and in 59 food types (Figures S2E and S2F). Surprisingly, 49% of the SGBs found in dairy lacked any reference genomes despite being the most studied category in food microbiome studies (Figure S2E), and comparable percentages were obtained in fermented beverages (42%) and fermented seeds (45%). The least characterized food types accordingly to the per-sample ratio of uSGBs reconstructed over kSGBs were Mexican pulque (median ratio = 5 from 5 samples), African palm wine (median ratio = 2, 6 samples), and Korean skate fish (median ratio = 1, 5 samples) (Figures S2A and S2B). Even more consumed food types had a non-negligible uSGB content such as post-processed vegetables (median ratio = 0.63, 6 samples), pre-processed vegetables (median ratio = 0.6, samples = 6), coffee (median ratio = 0.5, samples = 6), and sauerkraut (median ratio = 0.5, samples = 9). Cheese brine was the type in dairy with the highest uSGB fraction (53 samples), while similar distributions were exhibited by large fermented types such as cheese (1,043), kefir (284), and non-fermented (raw) milk (n = 110). Overall, these results underscore that almost the entire range of food types tested in this study appear to carry a blind spot concerning microbial characterization that should be further investigated for food development and production.

When considering the taxonomic structure of these uSGBs, the phylum Actinomycetota had the highest number of contributing MAGs (Figures S2C and S2D): 927 MAGs from 131 uSGBs, including 12 of the 20 most prevalent uSGBs. These prevalent uSGBs mainly belonged to the genera *Brevibacterium*, *Garicola*, and *Ruania*, as well as several genera comprising only uSGBs (from dairy, fermented beverages, and fermented meat; Figure 3). The phylum Bacillota comprised 710 MAGs from 175 uSGBs (65 uSGBs from the order Lactobacillales; Table S2) that spanned all categories: mainly from dairy (63% of the MAGs overall), although fermented fish (6.2 Bacillota MAGs/sample on average), fermented seeds (1.2), and fermented tubers and roots (0.8) were the most represented categories when normalizing for sample size. The phylum Pseudomonadota exhibited comparable numbers (407 MAGs from 150 uSGBs) and was found in ten categories, especially from cheese and cheese brine (244 MAGs). The reconstructed MAGs from water kefir (19 MAGs) and milk kefir (38) enabled the expansion of 11 uSGBs (for a total of 67 MAGs) of the genus *Acetobacter*, suggesting the presence of many yet-uncharacterized AABs in these food types.

(B) Number of food MAGs for each SGB detected in food.
(C) Distribution in the number of food MAGs according to the SGB unknownness level: known SGB (kSGB), unknown SGB (uSGB), and unknown food-specific SGB (ufSGB). Additional distributions are shown in Figures S2C–S2F.
(D) Number of SGBs in function of the food-human MAGs ratio defined as the ratio between the number of food MAGs and the number of food + human MAGs included in a specific SGB.
See also Figures S1 and S2.

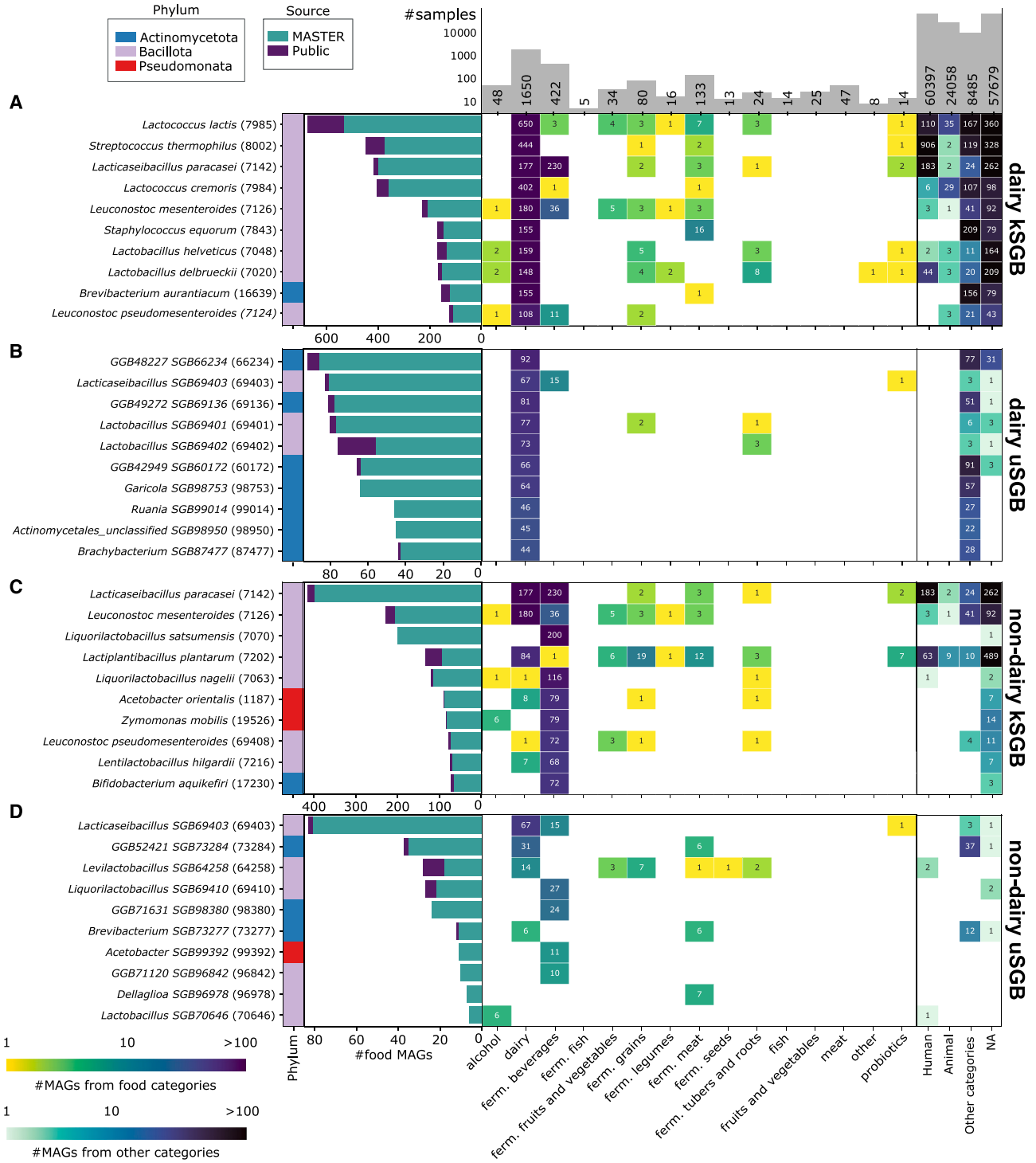


Figure 3. Distribution of the prokaryotic MAGs across food categories

The ten SGBs with the highest number of MAGs for (A–C) kSGBs (SGBs with at least one reference genome) and (B–D) uSGBs (SGBs comprising only MAGs) from (A and B) dairy and (C and D) non-dairy sources. In each panel, SGBs are ordered based on the number of food MAGs. The number of food MAGs is reported overall and for each of the 15 categories, along with the number of MAGs available in our MetaRefSGB repository and retrieved from other sources. The same representation for ufSGBs is reported in Figure S2.

See also Figure S2.

Overall, the SGB66234 was the most prevalent uSGB (92 MAGS and phylogenetically placed in the family *Corynebacteriaceae*; Figures 3B and S1A), was detected in ten European cheese types from seven countries, and was not identified in either human or animal metagenomes. Also, SGB69136 (81 MAGs and phylogenetically placed in the family *Microbacteriaceae*) was cheese-specific (from eight cheese types and six countries), was retrieved often from the same samples (69 Austrian cheese samples in common with SGB66234), and was similarly absent from human and animal metagenomes. Other non-dairy-specific uSGBs included *Lacticaseibacillus* SGB69403 (83 MAGs from cheese and water kefir) as well as *Lactobacillus* SGB69401 (80 MAGs) and *Lactobacillus* SGB69402 (76 MAGs) both found in dairy and occasionally in Nigerian fermented grains and fermented tubers and roots.

The consistent detection of yet-to-be-isolated microbial species associated with food microbiomes highlights their complexity and the need to perform targeted cultivation-based investigations to harness species and strains of potential technological relevance.

More than half of unknown food species are not detected in other environments

We next focused on those uSGBs detected exclusively in food sources and without matches in any of the other >1 M MAGs from other environments (mainly from human, animal, soil, water, and plants; Figures 2D, 3, and S2). We named them unknown food-specific SGBs (ufSGBs), and they represent the most understudied food-associated species to be targeted in future studies. More than half uSGBs (n = 290; 58%) were labeled as ufSGBs, which comprised 534 MAGs from 327 samples (Figure 2B) and were mainly associated with the phyla Bacillota (121 ufSGBs), Pseudomonadota (88), Actinomycetota (53), and Bacteroidota (16) (Figure S2C). These ufSGBs spanned all categories, especially fermented fish (5.8 MAGs from ufSGBs per sample on average), fruits and vegetables (1.3), and fermented seeds (1.2) when normalizing for sample size. They originated from 43 types, mainly cheese (74 ufSGBs), cheese brine (45), water kefir (34), post-processed vegetables (19), and skate fish (18).

The most reconstructed ufSGBs were ufSGB98380 (Figure S2G) with 24 MAGs from water kefir of diverse sources. It belonged to a monophyletic subtree comprising three other ufSGBs reconstructed from water kefir (i.e., ufSGB98379, ufSGB98381, and ufSGB98382; Figure S1A) and placed phylogenetically close to the family *Bifidobacteriaceae*. Also, ufSGB96887 (15 MAGs and assigned to the genus *Lactococcus*) was often retrieved from different dairy sources (i.e., raw milk, cheese brine, whey, and cheese products) from Southern Italy. Moreover, ufSGB99143 comprised 15 MAGs from Austrian Alpine cheeses and belonged to the genus *Marinobacter* that comprises halophilic bacteria.[92]

In general, ufSGBs were specific to single categories (98% of total ufSGBs) and types (93%), although six of them were detected in two categories (Table S2): ufSGB92515 (retrieved from dairy and fermented beverages, assigned to the genus *Hafnia*; Figure S2G); ufSGB94441 (dairy and fermented beverages; genus *Lactiplantibacillus*); ufSGB94442 (fermented fruits and vegetables and fermented meat; genus *Lactiplantibacillus*); ufSGB96932 (fermented seeds and fermented tubers and roots; genus *Atopostipes*); ufSGB94707 (dairy and fermented fruits and vegetables; family Oceanospirillaceae); and ufSGB96974 (alcohol and fermented tubers and roots; phylum Bacillota).

The high proportion of uncharacterized microbes found uniquely in food represents an opportunity for future research aimed at characterizing these species and their contribution to the features of the corresponding food products.

Food categories have distinctive quantitative microbial traits

Until now, our results were dependent on an assembly based approach that generated MAGs. We further performed a more sensitive and quantitative taxonomic profiling through MetaPhlAn 4, whose database[84] was expanded to include the SGBs defined in this work (STAR Methods). Of the resulting 29,969 SGBs that can be comprehensively detected at coverage as low as 0.1×, 3,622 SGBs were identified in at least one food metagenome (Figure 1E; Table S2).

The within-sample alpha diversity was highly variable across metagenomes (mean ± SD: 25 ± 30 for Richness and 1.2 ± 0.7 for Shannon index) and food categories. Non-fermented fish had the highest Shannon index (median = 2.3, Figures 4A and S3G), followed by fermented tubers and roots (2.1), meat (2.0), fermented seeds (2.0), fermented legumes (1.9), and fruits and vegetables (1.8), with results broadly paralleled by the estimated richness (Figures S3A and S3H). When comparing non-fermented with fermented food categories (i.e., meat with fermented meat, fish with fermented fish, and fruits and vegetables with fermented fruits and vegetables), the non-fermented foods exhibited a higher microbial diversity (Figure 4A), reflecting the selective pressure of fermentation processes. Large variability was observed in the largest food category, dairy (mean ± SD: 1.0 ± 0.7 for Shannon index and 24 ± 27 for richness; Figures 4A and S3A), with alpha diversity measures linked to food types (Figure S3D). The highest diversities were found in cheese brine (median Shannon index = 1.3, richness = 38), wara (Shannon index = 1.4, richness = 17), and cheese (Shannon index = 0.9 and richness = 21); for the latter one, samples before ripening showed slightly lower diversity than the final product, and similar values were found in non-fermented (raw) milk. Unknown species constituted a fundamental fraction: 48% of the samples (1,173) contained at least one uSGB or ufSGB. We further computed the ratio between the number of uSGBs and kSGBs (Figures S3E and S3F), and 605 samples (24%) had a ratio >0.1. Consistently with assembly results, pulque and fermented tea (e.g., kombucha and pu-erh) were the types carrying the higher unknown fraction. For dairy, cheese brine was the most uncharacterized, and cheese before ripening was more characterized than final products (Figure S3F).

Beta diversity distributions were driven by the food category (permutational multivariate analysis of variance) PERMANOVA R^2 = 0.15, p < 1e–3; Figures 4B and S3C), and all comparisons between categories resulted in statistically significant differences (Figure S3I). Not surprisingly, the starting raw material was the main distinguishing factor as it carries the raw material microbiota and shares other features (such as nutrient content and pH) that influence the final microbial composition. This

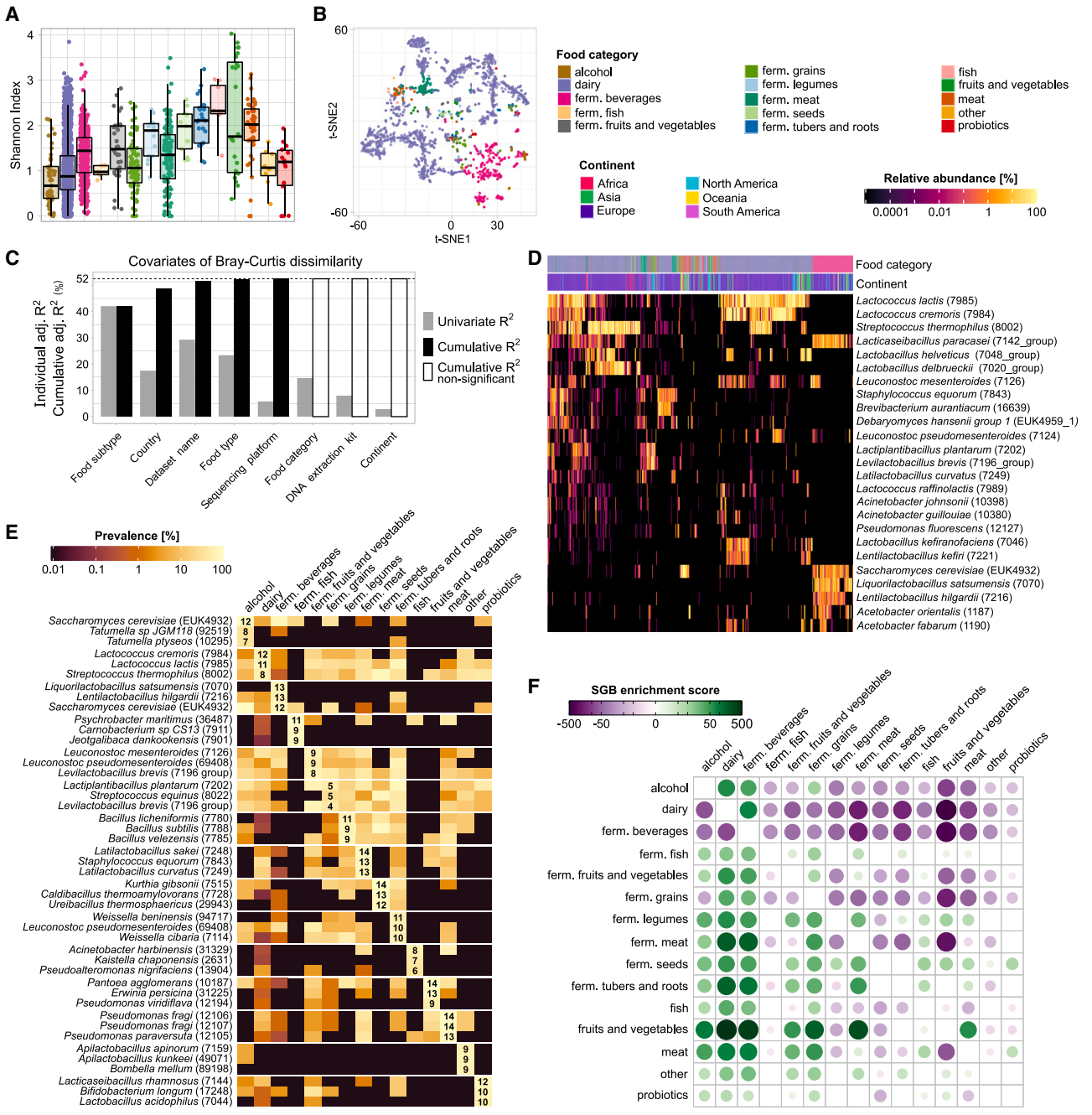


Figure 4. Taxonomic profiling enables sensitive characterization of food metagenomes

(A and B) Differences in terms of (A) alpha (Shannon index; Richness in Figure S3A) and (B) beta (t-distributed stochastic neighbor embedding [t-SNE] dimensionality reduction using Bray-Curtis distances) diversity for the food metagenomes grouped into 15 categories.

(C) Permutation tests in constrained ordination on distance-based redundancy analysis: individual (left bars) and cumulative (right bars) contributions for each variable based on Bray-Curtis dissimilarity.

(D) Relative abundance (rel. ab.) from taxonomic profiles for the 25 SGBs most prevalent in food, along with information on food category and continent of origin.

(E) The most representative SGBs for each food category along with their prevalence across categories. Numbers represent the statistically significant comparisons between categories. Other differentially prevalent SGBs are reported in Figure S4B, and the same representation specific to uSGBs is shown in Figure S4C.

(F) The SGB enrichment score (see STAR Methods) for each pair of food categories. A score > 0 indicates a higher number of SGBs enriched in the row category. See also Figures S3 and S4.

was confirmed by the contribution of variables to beta diversity (Figure 4C) with a total cumulative adjusted R^2 of 52% mainly explained by food subtype (univariate adj. R^2 = 42%), dataset name (29%), food type (23%), and country (17%). The identification of the dataset name as a covariate could imply technical biases across datasets, but it is more likely the effect of different studies covering different and specific categories and types. For dairy, cheese before and after ripening overlapped in the ordination plot, while more distinct clusters were associated with non-fermented (raw) milk and cheese brine.

We also assessed the predictability of the sample category according to taxonomic profiles by leveraging machine learning-based predictive modeling (STAR Methods). For one-vs.-all category comparisons, the area under the ROC curve (AUC) was always close to 1 (mean ± SD = 0.97 ± 0.06, median = 0.99, Table S3). The accuracy remained high when comparing more similar foods; in dairy, the sample type was predicted with AUC = 0.97 ± 0.05 (Table S3). Similar predictability levels were obtained even when looking at the finest food categorization level, and 23 commercial cheese types were discriminated with AUC = 0.97 ± 0.04 (Table S3; STAR Methods). These results open interesting perspectives for microbiome-based quality-control strategies in the food system, supporting the future application of metagenomics in food traceability and authentication.

Highly prevalent species determine subgrouping within food categories

The clustering of samples according to food categories was also evident by considering the 25 most prevalent SGBs alone (Figure 4D). Some level of clustering was also associated with geography (both at continent and country level), but this was due to the correlations between categories/types and geographic origin and should not be interpreted as generalizable notions.

We further identified subgrouping within specific categories. For example, dairy clustered into multiple subgroups (Figure 4D; Table S2): the largest group was dominated by *L. lactis* and *L. cremoris*, which tended to co-occur (e.g., in Dutch-type and blue cheeses); another subgroup was characterized by high occurrences of *S. thermophilus* coupled with *L. paracasei* (e.g., in Fontina cheese) or by *Lactobacillus helveticus* coupled with *L. delbrueckii* (e.g., in mozzarella). *L. kefiranofaciens* and *Lentilactobacillus kefiri*[93,94] co-occurred in milk kefir, while water kefir defined a tight cluster dominated by *L. paracasei*, *L. satsumensis* (not detected in any other types), and *Lentilactobacillus hilgardii*. Other fermented beverage types (e.g., kombucha, coffee, and pu-erh tea) clustered instead more closely with alcoholic beverages samples.

The set of 25 most prevalent SGBs also included the two eukaryotic species *Saccharomyces cerevisiae* and *Debaryomyces hansenii* group 1, highlighting the importance of fungi, yeasts in particular, when exploring the microbial diversity of food microbiomes.

Food-specific microbial signatures encompass known and unknown species

We determined category-specific microbial signatures by assessing differentially abundant SGBs and computing the specificity value as the number of one-vs.-one significant comparisons (Figures 4E and S4B; Table S3; STAR Methods). Looking at specific categories, fermented meat, for instance, was characterized by *L. sakei* (typically present during ripening, aging, and fermentation), *S. equorum*, and *Latilactobacillus curvatus*.[58,95–98] Non-fermented meat, instead, was enriched in three distinct SGBs of *Pseudomonas fragi* (i.e., SGB12106 corresponding to *Pseudomonas_E bubulae* in GTDB,[91] SGB12107—*Pseudomonas_E fragi*, and SGB12108—*Pseudomonas_E fragi_D*), a species associated with dairy and raw meat spoilage.[99,100] Fermented legumes and fermented seeds were characterized by several SGBs from the genus *Bacillus*, which has important roles in food (particularly soybean) fermentation.[101,102] We also defined the SGB enrichment score (STAR Methods) and identified fermented fish, fermented legumes, fermented seeds, fermented tubers and roots, and meat as the categories with the highest number of category-specific enriched SGBs (Figure 4F); this aligned roughly with the categories having lower intra-category diversity (Figure S3B). In dairy, similar numbers were obtained when non-fermented samples were disentangled from fermented ones (Figure S4D). Such findings suggested that not only whole microbial communities but also several food-specific SGBs can be considered as markers of food categories and even types, which could be harnessed in food traceability and authentication, as mentioned above.

We also looked specifically at uSGBs to assess the contribution of uncharacterized species for the specificity of food category microbiomes and identified 81 uSGBs significant in at least one comparison (Figure S4C). The highest specificity was obtained by *Liquorilactobacillus* SGB69410 in fermented beverages, a category that included ten additional discriminative uSGBs. Other categories enriched in uSGBs were fruits and vegetables and meat, in both fermented and non-fermented forms (Figure S4C). These results provided further evidence of the large reservoir of yet-to-be-isolated species in food microbiomes that could be of relevance in several applications, including food control.

Food and human microbial species overlap more in infants than adults

We extended the analysis to human metagenomes to test whether food-associated bacteria were among the prevalent colonizers of the human microbiome. We considered 19,833 gut and oral human samples from 39 countries available in cMD[27] (Figure S5A; Table S4; STAR Methods) and found 1,409 SGBs detected at least once in both food and human environments (Figure 5A; Table S5). We identified 816 food-prevalent SGBs (i.e., detected in $\geq$4 food samples with rel. ab. > 0.1%, see STAR Methods), and 409 of them were also detected in human samples (mean ± SD 5 ± 4 food SGBs per human sample; Figure 5A).

Several host conditions impacted the amount of food species found in the human microbiome. Food microbiomes overlapped with stool more than oral microbiomes across all age categories (Figures 5B and S5B, p < 1e–100). The number of SGBs in common between food and human microbiomes was higher in Westernized (W) compared with non-Westernized (NW) populations (Figure 5B, p < 1e–60), which however possibly reflected

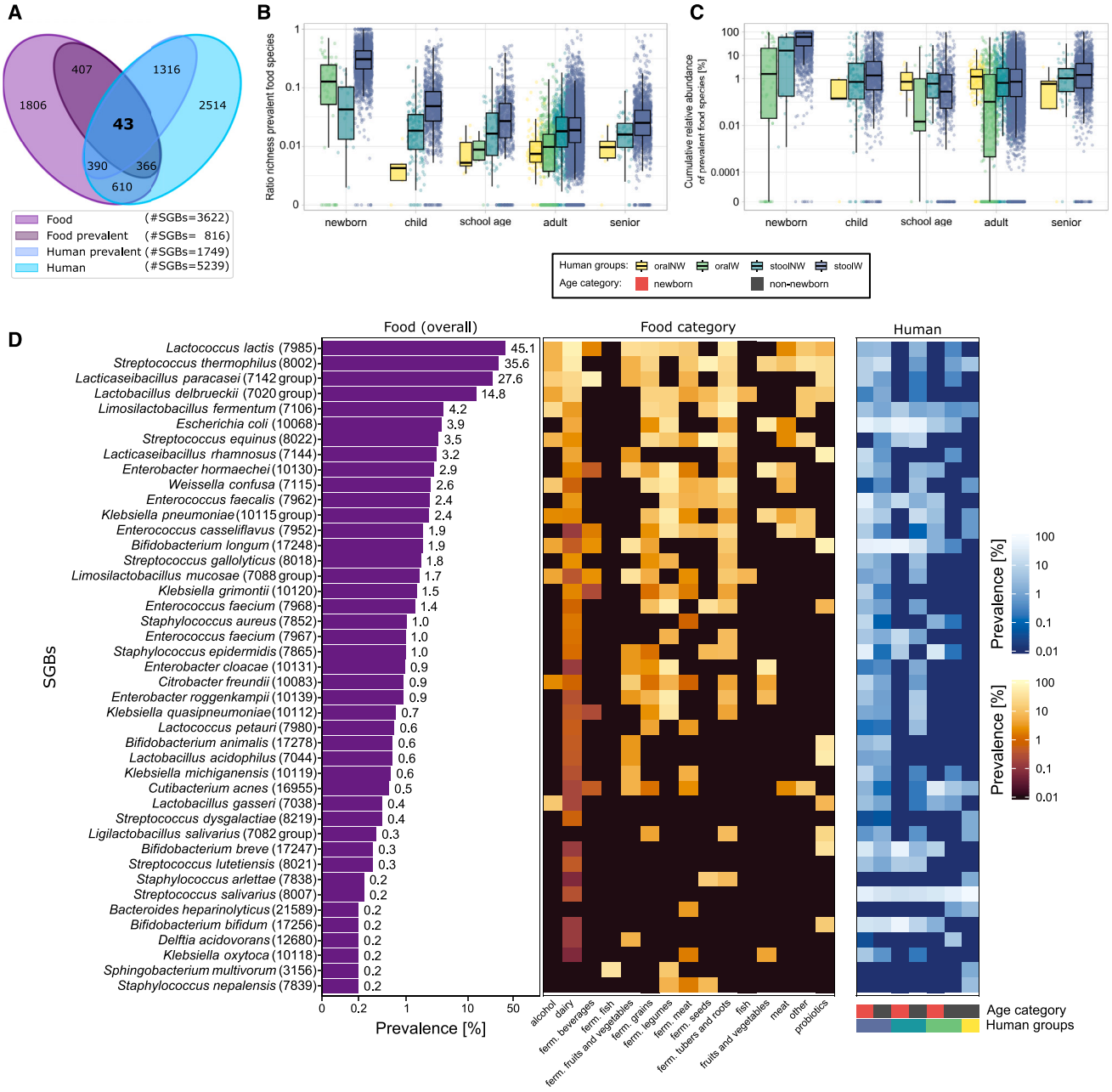


Figure 5. Overlap of SGBs prevalent in both food and human microbiomes

(A) Number of SGBs detected in food (n = 2,533) and human (n = 19,833) metagenomes from taxonomic profiles. We identified 409 SGBs as prevalent in food samples (i.e., found in ≥4 food samples with rel. ab. ≥ 0.1%) and detected at least once in human microbiomes.

(B and C) (B) The number of these 409 food-prevalent SGBs detected in human samples and normalized by each sample total richness and stratified by multiple host characteristics (i.e., age category, body site, and lifestyle). Similarly, cumulative rel. ab. in human metagenomes of these food-prevalent SGBs is reported in (C).

(D) Out of these 409 SGBs, 43 of them were identified as prevalent in humans, i.e., found in ≥1% of the samples with rel. ab. ≥ 0.1% in at least one human subgroup: stoolW (stool from Westernized populations, n = 17,884); stoolNW (stool from non-Westernized populations, n = 1,092); oralW (oral from Westernized populations, n = 694); and oralNW (oral from non-Westernized populations, n = 163). We show prevalence of these 43 SGBs in food metagenomes overall and in each food category, along with their prevalence across the human sub-groups and age categories (rel. ab. was thresholded at 0.1% to identify positive samples). Summary statistics are reported in Table S6.

See also Figure S5.

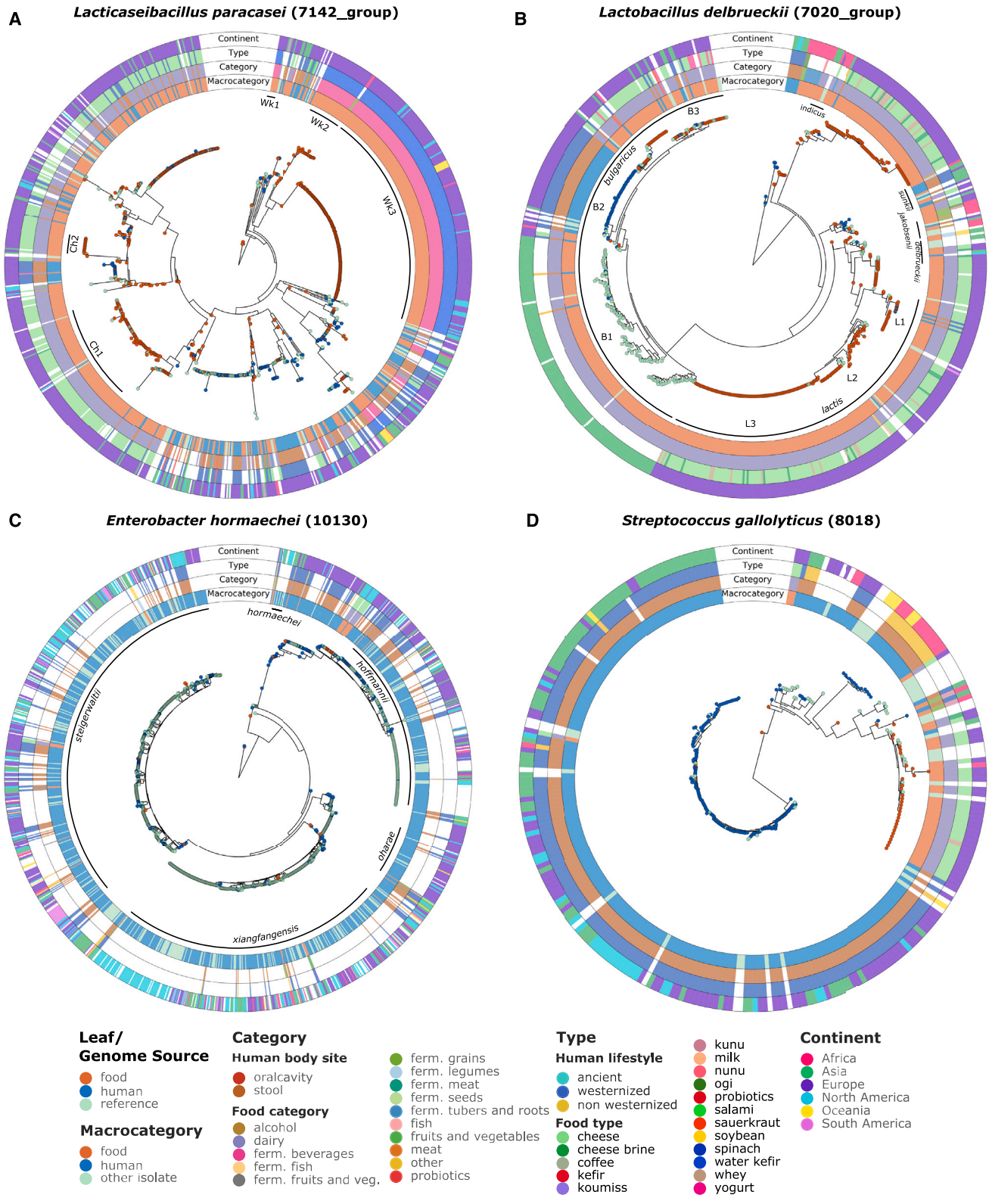


Figure 6. Phylogenetic trees of relevant prokaryotic SGBs highlight strain-level overlaps between food and human sources

Trees generated using StrainPhlAn; clades reported in literature or food-specific are annotated.

(A) *L. paracasei*: Wk1, Wk2, and Wk3 with water kefir strains; *Ch1* and *Ch2* with European cheese strains.

(legend continued on next page)

sampling biases. Within the same body site and across age categories and lifestyle, the absolute number of food SGBs tended to remain stable (Figures S5C and S5D), but their contribution when normalizing by the sample richness (Figures 5B, S5E, and S5F) and when considering their cumulative rel. ab. (Figure 5C) was considerably higher in newborns (mean cumulative rel. ab. = 56%) and children (8%) compared with school ages (3%), adults (3%), and seniors (5%).

Some food species were detected in humans consistently across age categories: *Bifidobacterium longum*, *Escherichia coli*, *Streptococcus salivarius*, *S. thermophilus*, *Bifidobacterium bifidum*, and *Bifidobacterium breve* (Figure 5D). While it is expected that these species, especially those from the genus *Bifidobacterium*, are found in newborns, their detection in adults when diets become more diverse is noteworthy. Although it is unlikely that infants acquire many of such species directly from food consumption since maternal transmission is the most probable route of acquisition,[103–105] food could nonetheless be seeders of adult microbiomes whose strains are further transmitted person-to-person and possibly retained into adulthood.

We further focused on non-rare human microbiome members (rel. ab. > 0.1% in >1% of samples; STAR Methods) and identified 43 SGBs prevalent in both food and humans (Figures 5A and 5D). Unsurprisingly, 21 of them belonged to LAB (Figure 5D; Table S6), which play a vital role in dairy—the most sampled category—and are commonly found in the human gut.[45,87] *L. lactis* and *L. fermentum* were widespread across lifestyles, *L. paracasei* and *L. delbrueckii* were more prevalent in W, and *S. thermophilus* was prevalent in stool from W (16% with rel. ab. > 0.1%) and oral cavity from NW (4%). Ten SGBs belonged to the family *Enterobacteriaceae*, with *E. coli* being the most prevalent. Finally, four shared SGBs belonged to family *Bifidobacteriaceae*; *B. longum* had the highest prevalence in stool (50% in W and 25% in NW), while *Bifidobacterium animalis* was detected in stool exclusively from W (4% prevalence).

Overall, several species found in food were also identified in the human microbiome, with a large number of stool and oral samples containing food species. Such overlapping species frequently accounted for a large fraction of the infant microbiome, while its rel. ab. was lower in adults (mean ± SD = 3% ± 7%; Figure 5C).

Identification of common strains between food and human microbiomes

We further performed strain-level and strain-matching analysis via StrainPhlAn 4[84] (STAR Methods) and identified potential transmission patterns for some food-human overlapping SGBs (Figures 6 and S6).

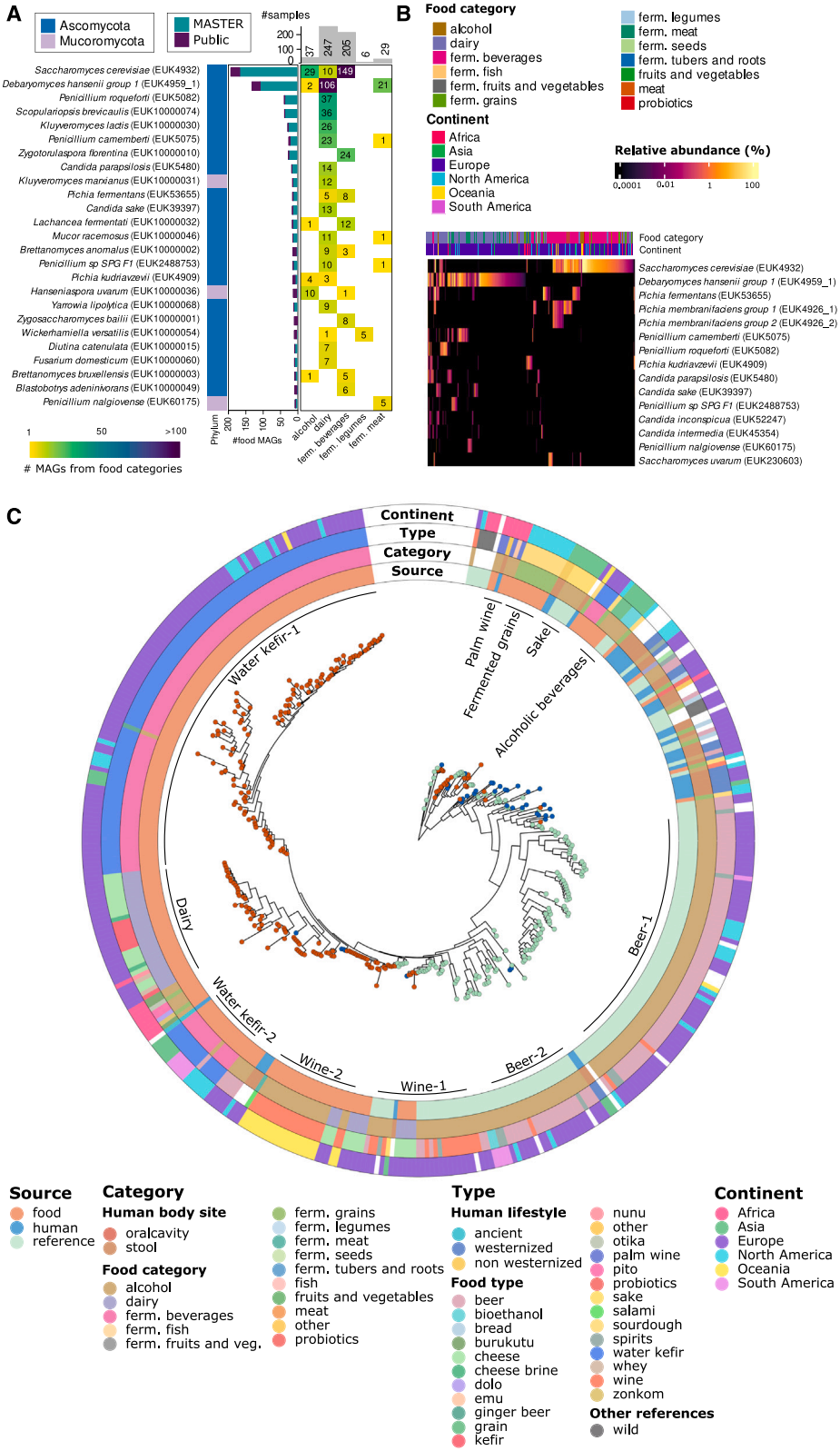
For example, *L. paracasei* was prevalent in dairy (33%), fermented beverages (74%), and water kefir (79%), and its strains were spread with the human ones across the phylogeny (Figure 6A). Nevertheless, water kefir strains clustered into three phylogenetically close subtrees regardless of geographic origin (Wk1, Wk2, and Wk3) that did not include any other food or human strains. Several European cheeses (n = 80) defined the clade Ch1, which comprised reference genomes from two stool samples and commercial and artisan dairy sources worldwide, suggesting a common industrial strain origin. Asturian cheeses along with a strain isolated from Asian fermented goat milk clustered in clade Ch2. Overall, human strains were spread across the whole phylogeny and were quite similar to the food ones (57% of human MAGs having ANI > 99.99% from at least one food MAG), pointing to food as the most probable source of the strains found in the human gut.

L. delbrueckii, widely employed in the dairy industry, exhibited six main subspecies (subsp.) in accordance with literature[106] (Figure 6B). The subsp. *bulgaricus*, relevant for yogurt production, was the most reconstructed from both food and human samples. This subsp. comprised three main clusters: different Asian dairy products along with two Asian human strains in B1; Dutch human strains in B2; yogurt, Italian cheeses, and six human strains in B3. We identified three clusters also for subsp. *lactis*, common in European cheese: human strains in L1, Italian cheeses in L2, and Austrian Alpine cheese in L3 strains. In addition to dairy, several strains were instead retrieved from African fermented tubers and roots and fermented grains and mainly assigned to subsp. *indicus* and subsp. *jakobsenii.*

Besides characterizing species typically involved in food processing, we also investigated potential pathogens. Most species of known concern for foodborne transmission were rarely found in our sampled foods (e.g., n = 1 for *Listeria monocytogenes* and n = 3 for *Clostridium perfringens*). Other species with potential pathogenic strains were more frequent (e.g., n = 95 with rel. ab = 0.01% for *Staphylococcus aureus*, Figure S6L; n = 173 with rel. ab. = 0.02% for *E. coli*, Figure S6C). The ESKAPEE species *E. hormaechei*[107] was detected in 77 food samples with rel. ab. = 0.02% (Figure 6C) warranting further investigation for food safety purposes.[108–110] The phylogeny included 1,023 reference genomes from humans as well as other animals, food, and natural environments and 38 food strains (n = 19 from cheese, 7 from American spinach,[62] and 5 from Asian fermented soybeans), and we detected at least 5 subspecies in accordance with literature[111] (all but subsp. *oharae* found in humans and food).

Streptococcus gallolyticus is similarly potentially associated with disease, including colorectal cancer[112–114] (11% prevalence vs. 3% in healthy controls; Table S4), and was also prevalent in food. However, food strains clustered separately from the human ones (Figure 6D), thus likely excluding a food origin of the human strains. Metagenomic surveys of food samples can therefore be of relevance for investigating the source of origin of human microbiome species with known, as well as less-appreciated, pathogenic potential.

(B) *Lactobacillus delbrueckii*: L1, L2, and L3 associated with *subsp. lactis*; B1, B2, and B3 associated with *subsp. bulgaricus*.
(C) *Enterobacter hormaechei* included at least 5 subsp.: *hormaechei*, *hoffmannii*, *oharae*, *xiangfangensis*, and *steigerwaltii*.
(D) *Streptococcus gallolyticus*. Isolation sources for the reference genomes are summarized in Table S7. Additional trees for relevant SGBs are reported in Figure S6.
See also Figure S6.



(legend on next page)

Fungal taxa are widespread members of food metagenomes

After investigating the distribution of prokaryotes in food microbiomes, we set out to characterize its eukaryotic fraction. We reconstructed 787 MAGs of eukaryotic origin and of sufficient quality for downstream analysis (n = 392 HQ and n = 395 MQ; STAR Methods; Table S2). These MAGs were recovered from dairy (401 MAGs from 247 samples), fermented beverages (233 MAGs from 205 samples), and eight additional categories (i.e., alcohol, probiotics, and different fermented products; 153 MAGs from 91 samples; Figure 7A). These MAGs were clustered into 108 SGBs, all taxonomically assigned to fungal clades, and resulting in 78 kSGBs (from 742 MAGs) and 30 uSGBs (from 45 MAGs; Table S2). The most reconstructed uSGBs comprised 4 MAGs each and were assigned to families Aspergillaceae (EUK10000018), Mucoraceae (EUK10000045), and Aspergillaceae (EUK10000063).

S. cerevisiae was the most reconstructed SGB (191 MAGs), mostly recovered from fermented beverages (149) and alcoholic beverages (29). Other commonly reconstructed species were *D. hansenii* group 1 (130 MAGs mostly from dairy and fermented meat), *Penicillium roqueforti* (37 MAGs from multiple blue cheeses), *Scopulariopsis brevicaulis* (36 MAGs), and *Kluyveromyces lactis* (26 MAGs). Moreover, species such as *Pichia kudriavzevii* and *Wickerhamiella versatilis* were recovered from multiple food categories (Table S2).

The eukaryotic SGBs were detected in 45% of food samples according to taxonomic profiling (median cumulative rel. ab. = 1% when present; Figure 7B; Table S3). The highest prevalences were found in alcoholic beverages (97% prevalence), fermented beverages (92%), and fermented meat (68%), food categories characterized by high numbers of retrieved MAGs (Figure 7A). Nevertheless, profiling detected eukaryotes in other three categories (Figure 7B): fermented fish (20% prevalence), fruits and vegetables (12%), and meat (9%).

We identified some clusters according to category based on the most prevalent species (Figure 7B). These were mainly driven by the high prevalence of *S. cerevisiae* in alcoholic beverages (91%) and fermented beverages (76%), and the high prevalence of *D. hansenii* group 1 in fermented meat (64%) and dairy (24%). Specifically, *S. cerevisiae* was detected at an overall prevalence of 18% (mean rel. ab. = 9.9% when present) in 8 food categories, reflecting its widespread use as industrial yeast.[116] *D. hansenii* group 1 was instead found in 19% of food samples (mean rel. ab. = 3.9% when present) and across 7 food categories: highly common in nature,[117] it was detected in fermented meat (often inoculated to increase aromatic profiles in sausages and dry-meats[118]), dairy (mainly in cheese brine and cheese, particularly cheddar), meat, fermented seeds, and fermented beverages (e.g., in 67% of pu-erh tea samples).

Strain-level profiling of *S. cerevisiae* enables integration of metagenomes with isolate genomes

Finally, we explored the genetic diversity of *S. cerevisiae*, the most abundant eukaryote in our food metagenomes and among the most important within the food industry.[119] We performed phylogenetic analysis through StrainPhlAn and integrated our set of food and human metagenomes with 157 isolate genomes from liquid state fermentation (LSF, e.g., from wine and beer) and solid-substrate fermentation (SSF, e.g., from bread and sake) with yeast[115] (STAR Methods).

Our phylogeny (Figure 7C) comprised 424 strains and showed a stratification by food type and geography in accordance with literature.[21,116,120] We recovered well-defined yeast clusters such as Sake/Asian within SSF and Dairy, Beer-1, and Beer-2 within LSF. Strains from other alcoholic beverages (e.g., spirits) were less clustered, in agreement with previous studies.[115,121] Most yeast strains from wine were either within a Wine plus Dairy cluster (Wine-1) or a distinct cluster mainly constituted of Australian spontaneous wine fermentation samples (Wine-2).[74] Although strains from some of our food groups were within known clusters (e.g., African fermentation strains group with the Dairy cluster), most of them formed distinct clusters. We identified a large cluster composed entirely of water kefir samples and no reference genomes. Samples from North and South American water kefir displayed geographical differentiation, while milk kefir strains were within the dairy cluster, indicating a clear substrate origin.

Some of the *S. cerevisiae* strains recovered from the human gut were within food clusters, suggesting a possible strain transmission via food consumption. For example, the (African) Palm wine cluster included a subject from the same continent, and Asian individuals were within the (Asian) Sake cluster, and a European individual laid in a European cheese cluster. Phylogenetic branch supports for these groups were strong (STAR Methods; Figure S7A), possibly indicating the existence of some geographically distributed food-human transmitted strains. A small number of human-associated strains clustered instead with laboratory controls and probiotic strains, suggesting possible strain acquisitions from probiotic consumption. However, some clades received poor statistical support (Figure S7A) and did not pass a strict quality control selection of the alignment sites (Figure S7B). This advocates that strain characterization of *S. cerevisiae* may be complicated by their low

Figure 7. Fungi are prevalent in food microbiomes and span multiple food categories

(A) We show the 25 eukaryotic SGBs with the highest number of MAGs recovered from food metagenomes. The number of food eukaryotic MAGs is reported in total and for each food category with ≥5 MAGs, along with the number of samples from which these MAGs were recovered.

(B) Taxonomic profiles for the 15 most prevalent eukaryotic SGBs in food metagenomes. Only the 1,079 food metagenomes containing at least one of these eukaryotic SGBs are shown.

(C) Strain-level profiling enables characterization of *S. cerevisiae*. Phylogenetic tree built by mapping raw-reads against *S. cerevisiae* markers, followed by phylogeny reconstruction (STAR Methods). The tool retrieved 267 strains from metagenomes, which were integrated with 157 isolate genomes.[115] Leaves are colored according to source, external rings encode metadata information, and branches are annotated as in Maixner et al.[115] Non-circular, manually curated versions, and bootstrap values are reported in Figure S7.

See also Figure S7.

abundance in human metagenomes and that the interpretation of their phylogeny should be taken with care.

In general, most of the food clusters that we identified were well supported (Figure S7A), consistent over data treatment (Figure S7B), and in accordance with the known phylogeny of yeast, which indicated the effectiveness of our approach in performing phylogenetic analysis of yeasts.

DISCUSSION

In this work, we developed and described cFMD, a resource resulting from the collection of thousands of newly sequenced and publicly available food metagenomes with standardized metadata and data products. The sequences added in cFMD that we made available through the MASTER EU Consortium largely expanded the characterization of food microbiomes for its size (~3× more metagenomes than previously available), quality of sequencing (~2× higher depth and ~4× more MAGs), and types of foods (Figure 1). The generation of 10,899 food MAGs and their integration with >1 M available MAGs and genomes widened the genomic information of species typically employed in food production (Figure 3) along with the identification of 320 yet-to-be-isolated species (Figures 2 and 7).

Integrative analysis with >20,000 human metagenomes revealed an overlap between food and human microbiomes of 1,409 SGBs (Figure 5A), with these species explaining on average 11% of a single human sample with a cumulative rel. ab. of 3% (Figure 5C), even though an overlap in species is not a proxy of strain overlap and/or transmission. While this overlap is arguably moderate if compared with person-to-person species or even strain overlap (reaching ~50% strain overlap between mothers and infants), it still represents a sizable fraction of the human microbiome and also an important evolutionary pattern through which the human microbiome might have developed over the millenia. A large fraction of the infant gut microbiome contained food SGBs (mean = 56%, Figure 5C), despite direct acquisition from food being only one of the possible routes of transmission (e.g., vertical and horizontal transmission[42,44,122]). In adults, food SGBs constituted 3% of the microbial community (Figures 5B and 5C). Strain-level analysis also inferred potentially recent food-to-microbiome transmission events (Figure 6) also for eukaryotes such as *S. cerevisiae* (Figure 7C).

We detected hundreds of uncharacterized species in the food microbiome, opening new venues for their in-depth characterization. Indeed, targeted isolation and functional characterization of these dietary microbes we detected only by metagenomics should be the next step to further exploit their role in food processing, quality, and safety. However, the global diversity of dietary microbes is still far from unraveled, and as such, our resource should be the starting point for additional integrative initiatives on food microbiomes. For instance, our findings support molecular typing means to develop food authenticity and origin certifications based on microbial specificity (Figures 4D–4F), which would require additional expansion of the number of metagenomes of the same food types from different locations and industries.

The MASTER EU Consortium provided numerous sequences from multiple environments and sampling locations, spanning the food system from rumen to fish, factories, soils, fermented food, cereals, and vegetables. Integrating these resources would unlock several relevant applications: from the study of the microbiome evolution along the food system to the study of the diffusion of antimicrobial resistance or spoilage-related genes in foods, to the detection of pathogens in food quality control, and, finally, also to the study of transmission along the food-human axis.

Limitations of the study

Detecting direct food-to-human microbial transmission remains a challenging task. While we investigated possible transmission events (Figures 6, 7, S6, and S7), the timing cannot currently be estimated with precision and is hindered by concurrent person-to-person transmission. Even the directionality can be only inferred to some extent by the context, and whether there is a dose-dependent effect on the transmission remains an open question. Moreover, whether food microbes detected in the human gut are indeed microbiome colonizers or simply transient requires further investigation.[45] Intervention trials may be designed to supplement specific fermented foods to volunteers, while performing strain-level analysis of both foods and fecal samples (pre-, during, and post-administration) is important to investigate transmission events and strains stability and viability. Eukaryotic genome characterization from metagenomes is still underappreciated due to multiple limitations from possible biases in DNA extraction[123] to more challenging MAG reconstructions, and the approach employed can be a step toward standardizing these analyses.

The endeavor of a comprehensive collection of food microbiomes is far from complete, currently lacking examples from several countries and widely consumed food types. The current database composition is biased toward certain types of food (e.g., dairy) and geographic provenience. The inclusion of more diverse foods would favor the identification of unique microbial food markers, whereas the extension of sample metadata (e.g., processed vs. raw, ready-to-eat vs. cooked, ingredient information) could widen the possibility of applications and the value of the results. The currently defined and standardized metadata is based on previous work in the field,[81] further expanded by food microbiologists as part of the current work; nonetheless, the establishment of more general food ontologies for microbiome studies remains an ongoing process.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Overview of the resource
  - Survey of public datasets associated with food metagenomes
  - Collection of new food metagenomes
  - Integration of cFMD with human metagenomes from cMD
- METHOD DETAILS

- Metadata curation and standardization
- DNA isolation, library preparation, and sequencing of food samples
- Pre-processing of raw-reads
- Extraction of prokaryotic and eukaryotic MAGs
- Expansion of MetaRefSGB and ChocoPhlAn with the extracted MAGs
- Phylogenetic analysis for building the bacterial tree of life
- Generation of taxonomic and functional profiles

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Analysis of taxonomic profiles
  - Integrative analysis of food and human metagenomes
  - Strain-level characterization of common prevalent SGBs through StrainPhlAn

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.07.039.

ACKNOWLEDGMENTS

The MASTER EU Consortium was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no 818368. The work was also partially supported by the European Union's Horizon Europe programme (project DOMINO-101060218), the Italian Ministry of Foreign Affairs and International Cooperation (project FOODMICROHERITAGE-VN21GR09), the European Research Council (ERC-STG project MetaPG-716575 and ERC-CoG project microTOUCH-101045015) to N.S., the European Union's Horizon 2020 programme (projects ONCOBIOME-825410 and IHMCSA-964590) to N.S., the MUR PNRR INEST-Interconnected Nord-Est Innovation Ecosystem by NextGenerationEU (project ECS00000043) to N.S., the National Cancer Institute of the National Institutes of Health (projects 1U01CA230551 and U01CA261961) to N.S., the Premio Internazionale Lombardia e Ricerca 2019 to N.S, the European Union's Horizon 2020 programme (Marie Skłodowska-Curie project 101034371) to N.M.Q., the Spanish Ministry of Science and Innovation (Juan de la Cierva postdoctoral contract FJC2019-042125-I), the Science Foundation Ireland (SFI) under grant number SFI/12/RC/2273 (APC Microbiome Ireland), and SFI together with the Irish Department of Agriculture, Food, and the Marine, SFI/16/RC/3835 (VistaMilk) to P.D.C.

AUTHOR CONTRIBUTIONS

N.C., N.S., and E.P. conceived the study. C.B., S.B., R.C.R., I.C.-T., J.F.C.-D., F.D.F., J.L., S.K., D.O., N.M.Q., C.S., S.S., V.V., L.W., and V.T.M. collected the new food samples and extracted DNA. F.A. and E.S.K. generated metagenomic data. F. Asnicar and N.C. pre-processed the metagenomic reads. A.B.-M., M.P., N.C., and F. Asnicar contributed to the development of the bioinformatic pipeline. N.C. collected food metagenomes and curated the metadata. P.M. collected and curated the metadata of human samples. N.C. and E.P. performed the analysis. A.T., C.M., M.A., H.D., V.H., and O.R.S. contributed to the analysis. N.C. prepared the figures. F.P., M.C., N.S., and P.D.C. managed and coordinated this work inside the MASTER EU project. A.A.-O., A.M., V.T.M., M.W., D.E., P.D.C., and N.S. obtained the funding. N.C., E.P., and N.S. wrote the manuscript with contributions from P.D.C., D.E., A.T., O.R.S., G.F., and V.H. and input from all the authors. N.S. and E.P. supervised the work. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

D.O. is an employee of QIAGEN GmbH. E.S.K. is an employee of BaseClear B.V.

Received: January 12, 2024
Revised: May 17, 2024
Accepted: July 23, 2024
Published: October 3, 2024

REFERENCES

1. Jay, J.M., Loessner, M.J., and Golden, D.A. (2005). History of Microorganisms in Food. In Modern Food Microbiology (Springer), pp. 3–9.
2. Caplice, E., and Fitzgerald, G.F. (1999). Food fermentations: role of microorganisms in food production and preservation. Int. J. Food Microbiol. *50*, 131–149. https://doi.org/10.1016/s0168-1605(99)00082-3.
3. Dimidi, E., Cox, S.R., Rossi, M., and Whelan, K. (2019). Fermented Foods: Definitions and Characteristics, Impact on the Gut Microbiota and Effects on Gastrointestinal Health and Disease. Nutrients *11*, 1806. https://doi.org/10.3390/nu11081806.
4. Yap, M., Ercolini, D., Álvarez-Ordóñez, A., O'Toole, P.W., O'Sullivan, O., and Cotter, P.D. (2022). Next-Generation Food Research: Use of Meta-Omic Approaches for Characterizing Microbial Communities Along the Food Chain. Annu. Rev. Food Sci. Technol. *13*, 361–384. https://doi.org/10.1146/annurev-food-052720-010751.
5. De Filippis, F., Valentino, V., Alvarez-Ordóñez, A., Cotter, P.D., and Ercolini, D. (2021). Environmental microbiome mapping as a strategy to improve quality and safety in the food industry. Curr. Opin. Food Sci. *38*, 168–176. https://doi.org/10.1016/j.cofs.2020.11.012.
6. Marco, M.L., Heeney, D., Binda, S., Cifelli, C.J., Cotter, P.D., Foligné, B., Gänzle, M., Kort, R., Pasin, G., Pihlanto, A., et al. (2017). Health benefits of fermented foods: microbiota and beyond. Curr. Opin. Biotechnol. *44*, 94–102. https://doi.org/10.1016/j.copbio.2016.11.010.
7. Leeuwendaal, N.K., Stanton, C., O'Toole, P.W., and Beresford, T.P. (2022). Fermented Foods, Health and the Gut Microbiome. Nutrients *14*, 1527. https://doi.org/10.3390/nu14071527.
8. Tamang, J.P. (2014). Microfloras of Fermented Foods. In Encyclopedia of Food Microbiology, Second Edition, C.A. Batt and M.L. Tortorello, eds. (Academic Press), pp. 250–258.
9. Ercolini, D. (2004). PCR-DGGE fingerprinting: novel strategies for detection of microbes in food. J. Microbiol. Methods *56*, 297–314. https://doi.org/10.1016/j.mimet.2003.11.006.
10. Mayo, B., Rachid, C.T.C.C., Alegría, A., Leite, A.M.O., Peixoto, R.S., and Delgado, S. (2014). Impact of next generation sequencing techniques in food microbiology. Curr. Genomics *15*, 293–309. https://doi.org/10.2174/1389202915666140616233211.
11. Cocolin, L., and Ercolini, D. (2015). Zooming into food-associated microbial consortia: a "cultural" evolution. Curr. Opin. Food Sci. *2*, 43–50. https://doi.org/10.1016/j.cofs.2015.01.003.
12. De Filippis, F., Parente, E., and Ercolini, D. (2018). Recent Past, Present, and Future of the Food Microbiome. Annu. Rev. Food Sci. Technol. *9*, 589–608. https://doi.org/10.1146/annurev-food-030117-012312.
13. De Filippis, F., Pasolli, E., and Ercolini, D. (2020). The food-gut axis: lactic acid bacteria and their link to food, the gut microbiome and human health. FEMS Microbiol. Rev. *44*, 454–489. https://doi.org/10.1093/femsre/fuaa015.
14. Ferrocino, I., Rantsiou, K., and Cocolin, L. (2022). Microbiome and -omics application in food industry. Int. J. Food Microbiol. *377*, 109781. https://doi.org/10.1016/j.ijfoodmicro.2022.109781.
15. Ferrocino, I., Rantsiou, K., McClure, R., Kostic, T., de Souza, R.S.C., Lange, L., FitzGerald, J., Kriaa, A., Cotter, P., Maguin, E., et al. (2023). The need for an integrated multi-OMICs approach in microbiome science in the food system. Compr. Rev. Food Sci. Food Saf. *22*, 1082–1103. https://doi.org/10.1111/1541-4337.13103.
16. Ripp, F., Krombholz, C.F., Liu, Y., Weber, M., Schäfer, A., Schmidt, B., Köppel, R., and Hankeln, T. (2014). All-Food-Seq (AFS): a quantifiable screen for species in biological samples by deep DNA sequencing. BMC Genomics *15*, 639. https://doi.org/10.1186/1471-2164-15-639.
17. Walsh, A.M., Leech, J., Huttenhower, C., Delhomme-Nguyen, H., Crispie, F., Chervaux, C., and Cotter, P.D. (2023). Integrated molecular approaches for fermented food microbiome research. FEMS Microbiol. Rev. *47*, fuad001. https://doi.org/10.1093/femsre/fuad001.

18. Hellmann, S.L., Ripp, F., Bikar, S.-E., Schmidt, B., Köppel, R., and Hankeln, T. (2020). Identification and quantification of meat product ingredients by whole-genome metagenomics (All-Food-Seq). Eur. Food Res. Technol. *246*, 193–200. https://doi.org/10.1007/s00217-019-03404-y.

19. Verce, M., De Vuyst, L., and Weckx, S. (2019). Shotgun Metagenomics of a Water Kefir Fermentation Ecosystem Reveals a Novel Oenococcus Species. Front. Microbiol. *10*, 479. https://doi.org/10.3389/fmicb.2019.00479.

20. Wolfe, B.E., Button, J.E., Santarelli, M., and Dutton, R.J. (2014). Cheese rind communities provide tractable systems for in situ and in vitro studies of microbial diversity. Cell *158*, 422–433. https://doi.org/10.1016/j.cell.2014.05.041.

21. Díaz-Muñoz, C., Verce, M., De Vuyst, L., and Weckx, S. (2022). Phylogenomics of a Saccharomyces cerevisiae cocoa strain reveals adaptation to a West African fermented food population. iScience *25*, 105309. https://doi.org/10.1016/j.isci.2022.105309.

22. Yulandi, A., Suwanto, A., Waturangi, D.E., and Wahyudi, A.T. (2020). Shotgun metagenomic analysis reveals new insights into bacterial community profiles in tempeh. BMC Res. Notes *13*, 562. https://doi.org/10.1186/s13104-020-05406-6.

23. Yao, G., Yu, J., Hou, Q., Hui, W., Liu, W., Kwok, L.-Y., Menghe, B., Sun, T., Zhang, H., and Zhang, W. (2017). A Perspective Study of Koumiss Microbiome by Metagenomics Analysis Based on Single-Cell Amplification Technique. Front. Microbiol. *8*, 165. https://doi.org/10.3389/fmicb.2017.00165.

24. Arıkan, M., Mitchell, A.L., Finn, R.D., and Gürel, F. (2020). Microbial composition of Kombucha determined using amplicon sequencing and shotgun metagenomics. J. Food Sci. *85*, 455–464. https://doi.org/10.1111/1750-3841.14992.

25. Walsh, A.M., Macori, G., Kilcawley, K.N., and Cotter, P.D. (2020). Meta-analysis of cheese microbiomes highlights contributions to multiple aspects of quality. Nat. Food *1*, 500–510. https://doi.org/10.1038/s43016-020-0129-3.

26. Leech, J., Cabrera-Rubio, R., Walsh, A.M., Macori, G., Walsh, C.J., Barton, W., Finnegan, L., Crispie, F., O'Sullivan, O., Claesson, M.J., et al. (2020). Fermented-Food Metagenomics Reveals Substrate-Associated Differences in Taxonomy and Health-Associated and Antibiotic Resistance Determinants. mSystems *5*, e00522-20. https://doi.org/10.1128/mSystems.00522-20.

27. Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D.T., Beghini, F., Malik, F., Ramos, M., Dowd, J.B., et al. (2017). Accessible, curated metagenomic data through ExperimentHub. Nat. Methods *14*, 1023–1024. https://doi.org/10.1038/nmeth.4468.

28. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell *176*, 649–662.e20. https://doi.org/10.1016/j.cell.2019.01.001.

29. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., et al. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat. Biotechnol. *39*, 105–114. https://doi.org/10.1038/s41587-020-0603-3.

30. Muller, E., Algavi, Y.M., and Borenstein, E. (2022). The gut microbiome-metabolome dataset collection: a curated resource for integrative meta-analysis. NPJ Biofilms Microbiomes *8*, 79. https://doi.org/10.1038/s41522-022-00345-5.

31. Agostinetto, G., Bozzi, D., Porro, D., Casiraghi, M., Labra, M., and Bruno, A. (2022). SKIOME Project: a curated collection of skin microbiome datasets enriched with study-related metadata. Database (Oxford) *2022*, baac033. https://doi.org/10.1093/database/baac033.

32. Schmidt, T.S.B., Raes, J., and Bork, P. (2018). The Human Gut Microbiome: From Association to Modulation. Cell *172*, 1198–1215. https://doi.org/10.1016/j.cell.2018.02.044.

33. Singh, R.K., Chang, H.-W., Yan, D., Lee, K.M., Ucmak, D., Wong, K., Abrouk, M., Farahnik, B., Nakamura, M., Zhu, T.H., et al. (2017). Influence of diet on the gut microbiome and implications for human health. J. Transl. Med. *15*, 73. https://doi.org/10.1186/s12967-017-1175-y.

34. Bisanz, J.E., Upadhyay, V., Turnbaugh, J.A., Ly, K., and Turnbaugh, P.J. (2019). Meta-Analysis Reveals Reproducible Gut Microbiome Alterations in Response to a High-Fat Diet. Cell Host Microbe *26*, 265–272.e4. https://doi.org/10.1016/j.chom.2019.06.013.

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

36. Leshem, A., Segal, E., and Elinav, E. (2020). The Gut Microbiome and Individual-Specific Responses to Diet. mSystems *5*, e00665-20. https://doi.org/10.1128/mSystems.00665-20.

37. Wang, D.D., Nguyen, L.H., Li, Y., Yan, Y., Ma, W., Rinott, E., Ivey, K.L., Shai, I., Willett, W.C., Hu, F.B., et al. (2021). The gut microbiome modulates the protective association between a Mediterranean diet and cardiometabolic disease risk. Nat. Med. *27*, 333–343. https://doi.org/10.1038/s41591-020-01223-3.

38. Asnicar, F., Berry, S.E., Valdes, A.M., Nguyen, L.H., Piccinno, G., Drew, D.A., Leeming, E., Gibson, R., Le Roy, C., Khatib, H.A., et al. (2021). Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. Nat. Med. *27*, 321–332. https://doi.org/10.1038/s41591-020-01183-8.

39. Armet, A.M., Deehan, E.C., O'Sullivan, A.F., Mota, J.F., Field, C.J., Prado, C.M., Lucey, A.J., and Walter, J. (2022). Rethinking healthy eating in light of the gut microbiome. Cell Host Microbe *30*, 764–785. https://doi.org/10.1016/j.chom.2022.04.016.

40. Santonocito, S., Giudice, A., Polizzi, A., Troiano, G., Merlo, E.M., Sclafani, R., Grosso, G., and Isola, G. (2022). A Cross-Talk between Diet and the Oral Microbiome: Balance of Nutrition on Inflammation and Immune System's Response during Periodontitis. Nutrients *14*, 2426. https://doi.org/10.3390/nu14122426.

41. Valles-Colomer, M., Menni, C., Berry, S.E., Valdes, A.M., Spector, T.D., and Segata, N. (2023). Cardiometabolic health, diet and the gut microbiome: a meta-omics perspective. Nat. Med. *29*, 551–561. https://doi.org/10.1038/s41591-023-02260-4.

42. Valles-Colomer, M., Blanco-Míguez, A., Manghi, P., Asnicar, F., Dubois, L., Golzato, D., Armanini, F., Cumbo, F., Huang, K.D., Manara, S., et al. (2023). The person-to-person transmission landscape of the gut and oral microbiomes. Nature *614*, 125–135. https://doi.org/10.1038/s41586-022-05620-1.

43. Ercolini, D., and Fogliano, V. (2018). Food Design To Feed the Human Gut Microbiota. J. Agric. Food Chem. *66*, 3754–3758. https://doi.org/10.1021/acs.jafc.8b00456.

44. Manara, S., Selma-Royo, M., Huang, K.D., Asnicar, F., Armanini, F., Blanco-Miguez, A., Cumbo, F., Golzato, D., Manghi, P., Pinto, F., et al. (2023). Maternal and food microbial sources shape the infant microbiome of a rural Ethiopian population. Curr. Biol. *33*, 1939–1950.e4. https://doi.org/10.1016/j.cub.2023.04.011.

45. Pasolli, E., De Filippis, F., Mauriello, I.E., Cumbo, F., Walsh, A.M., Leech, J., Cotter, P.D., Segata, N., and Ercolini, D. (2020). Large-scale genome-wide analysis links lactic acid bacteria from food with the gut microbiome. Nat. Commun. *11*, 2610. https://doi.org/10.1038/s41467-020-16438-8.

46. Milani, C., Duranti, S., Napoli, S., Alessandri, G., Mancabelli, L., Anzalone, R., Longhi, G., Viappiani, A., Mangifesta, M., Lugli, G.A., et al. (2019). Colonization of the human gut by bovine bacteria present in Parmesan cheese. Nat. Commun. *10*, 1286. https://doi.org/10.1038/s41467-019-09303-w.

47. Tomofuji, Y., Kishikawa, T., Maeda, Y., Ogawa, K., Otake-Kasamoto, Y., Kawabata, S., Nii, T., Okuno, T., Oguro-Igashira, E., Kinoshita, M., et al. (2022). Prokaryotic and viral genomes recovered from 787 Japanese gut metagenomes revealed microbial features linked to diets, populations,

and diseases. Cell Genom. *2*, 100219. https://doi.org/10.1016/j.xgen.2022.100219.

48. Bertuzzi, A.S., Walsh, A.M., Sheehan, J.J., Cotter, P.D., Crispie, F., McSweeney, P.L.H., Kilcawley, K.N., and Rea, M.C. (2018). Omics-Based Insights into Flavor Development and Microbial Succession within Surface-Ripened Cheese. mSystems *3*, e00211-17. https://doi.org/10.1128/mSystems.00211-17.
49. Chacón-Vargas, K., Torres, J., Giles-Gómez, M., Escalante, A., and Gibbons, J.G. (2020). Genomic profiling of bacterial and fungal communities and their predictive functionality during pulque fermentation by whole-genome shotgun sequencing. Sci. Rep. *10*, 15115. https://doi.org/10.1038/s41598-020-71864-4.
50. Crovadore, J., Gérard, F., Chablais, R., Cochard, B., Bergman Jensen, K.K., and Lefort, F. (2017). Deeper Insight in Beehives: Metagenomes of Royal Jelly, Pollen, and Honey from Lavender, Chestnut, and Fir Honeydew and Epiphytic and Endophytic Microbiota of Lavender and Rose Flowers. Genome Announc. *5*, e00425-17. https://doi.org/10.1128/genomeA.00425-17.
51. De Roos, J., Verce, M., Weckx, S., and De Vuyst, L. (2020). Temporal Shotgun Metagenomics Revealed the Potential Metabolic Capabilities of Specific Microorganisms During Lambic Beer Production. Front. Microbiol. *11*, 1692. https://doi.org/10.3389/fmicb.2020.01692.
52. Du, R., Wu, Q., and Xu, Y. (2020). Chinese Liquor Fermentation: Identification of Key Flavor-Producing Lactobacillus spp. by Quantitative Profiling with Indigenous Internal Standards. Appl. Environ. Microbiol. *86*, e00456-20. https://doi.org/10.1128/AEM.00456-20.
53. Duru, I.C., Laine, P., Andreevskaya, M., Paulin, L., Kananen, S., Tynkkynen, S., Auvinen, P., and Smolander, O.-P. (2018). Metagenomic and metatranscriptomic analysis of the microbial community in Swiss-type Maasdam cheese during ripening. Int. J. Food Microbiol. *281*, 10–22. https://doi.org/10.1016/j.ijfoodmicro.2018.05.017.
54. Einson, J.E., Rani, A., You, X., Rodriguez, A.A., Randell, C.L., Barnaba, T., Mammel, M.K., Kotewicz, M.L., Elkins, C.A., and Sela, D.A. (2018). A Vegetable Fermentation Facility Hosts Distinct Microbiomes Reflecting the Production Environment. Appl. Environ. Microbiol. *84*, e01680-18. https://doi.org/10.1128/AEM.01680-18.
55. Escobar-Zepeda, A., Sanchez-Flores, A., and Quirasco Baruch, M. (2016). Metagenomic analysis of a Mexican ripened cheese reveals a unique complex microbiota. Food Microbiol. *57*, 116–127. https://doi.org/10.1016/j.fm.2016.02.004.
56. Ferrocino, I., Bellio, A., Giordano, M., Macori, G., Romano, A., Rantsiou, K., Decastelli, L., and Cocolin, L. (2018). Shotgun Metagenomics and Volatilome Profile of the Microbiota of Fermented Sausages. Appl. Environ. Microbiol. *84*, e02120-17. https://doi.org/10.1128/AEM.02120-17.
57. Smukowski Heil, C., Burton, J.N., Liachko, I., Friedrich, A., Hanson, N.A., Morris, C.L., Schacherer, J., Shendure, J., Thomas, J.H., and Dunham, M.J. (2018). Identification of a novel interspecific hybrid yeast from a metagenomic spontaneously inoculated beer sample using Hi-C. Yeast *35*, 71–84. https://doi.org/10.1002/yea.3280.
58. Kastman, E.K., Kamelamela, N., Norville, J.W., Cosetta, C.M., Dutton, R.J., and Wolfe, B.E. (2016). Biotic Interactions Shape the Ecological Distributions of Staphylococcus Species. mBio *7*, e01157-16. https://doi.org/10.1128/mBio.01157-16.
59. Kawai, T., Sekizuka, T., Yahata, Y., Kuroda, M., Kumeda, Y., Iijima, Y., Kamata, Y., Sugita-Konishi, Y., and Ohnishi, T. (2012). Identification of Kudoa septempunctata as the causative agent of novel food poisoning outbreaks in Japan by consumption of Paralichthys olivaceus in raw fish. Clin. Infect. Dis. *54*, 1046–1052. https://doi.org/10.1093/cid/cir1040.
60. Kumar, J., Sharma, N., Kaushal, G., Samurailatpam, S., Sahoo, D., Rai, A.K., and Singh, S.P. (2019). Metagenomic Insights Into the Taxonomic and Functional Features of Kinema, a Traditional Fermented Soybean Product of Sikkim Himalaya. Front. Microbiol. *10*, 1744. https://doi.org/10.3389/fmicb.2019.01744.
61. Landis, E.A., Oliverio, A.M., McKenney, E.A., Nichols, L.M., Kfoury, N., Biango-Daniels, M., Shell, L.K., Madden, A.A., Shapiro, L., Sakunala, S., et al. (2021). The diversity and function of sourdough starter microbiomes. eLife *10*, e61644. https://doi.org/10.7554/eLife.61644.
62. Leonard, S.R., Mammel, M.K., Lacher, D.W., and Elkins, C.A. (2016). Strain-Level Discrimination of Shiga Toxin-Producing Escherichia coli in Spinach Using Metagenomic Sequencing. PLoS One *11*, e0167870. https://doi.org/10.1371/journal.pone.0167870.
63. Li, Z., Feng, C., Luo, X., Yao, H., Zhang, D., and Zhang, T. (2018). Revealing the influence of microbiota on the quality of Pu-erh tea during fermentation process by shotgun metagenomic and metabolomic analysis. Food Microbiol. *76*, 405–415. https://doi.org/10.1016/j.fm.2018.07.001.
64. Li, Z., Dong, L., Zhao, C., and Zhu, Y. (2020). Metagenomic insights into the changes in microbial community and antimicrobial resistance genes associated with different salt content of red pepper (Capsicum annuum L.) sauce. Food Microbiol. *85*, 103295. https://doi.org/10.1016/j.fm.2019.103295.
65. Lordan, R., Walsh, A.M., Crispie, F., Finnegan, L., Cotter, P.D., and Zabetakis, I. (2019). The effect of ovine milk fermentation on the antithrombotic properties of polar lipids. J. Funct. Foods *54*, 289–300. https://doi.org/10.1016/j.jff.2019.01.029.
66. McHugh, A.J., Feehily, C., Fenelon, M.A., Gleeson, D., Hill, C., and Cotter, P.D. (2020). Tracking the Dairy Microbiota from Farm Bulk Tank to Skimmed Milk Powder. mSystems *5*, e00226-20. https://doi.org/10.1128/mSystems.00226-20.
67. Patro, J.N., Ramachandran, P., Barnaba, T., Mammel, M.K., Lewis, J.L., and Elkins, C.A. (2016). Culture-Independent Metagenomic Surveillance of Commercially Available Probiotics with High-Throughput Next-Generation Sequencing. mSphere *1*, e00057-16. https://doi.org/10.1128/mSphere.00057-16.
68. Pfefer, T. (2016). Evaluation of enriched microflora of raw milk cheese spiked with E. coli O157:H7 and E. coli O103 using next-generation sequencing technology. In IAFP 2016 Annual Meeting (IAFP).
69. Porcellato, D., and Skeie, S.B. (2016). Bacterial dynamics and functional analysis of microbial metagenomes during ripening of Dutch-type cheese. Int. Dairy J. *61*, 182–188. https://doi.org/10.1016/j.idairyj.2016.05.005.
70. Pothakos, V., De Vuyst, L., Zhang, S.J., De Bruyn, F., Verce, M., Torres, J., Callanan, M., Moccand, C., and Weckx, S. (2020). Temporal shotgun metagenomics of an Ecuadorian coffee fermentation process highlights the predominance of lactic acid bacteria. Curr. Res. Biotechnol. *2*, 1–15. https://doi.org/10.1016/j.crbiot.2020.02.001.
71. Quigley, L., O'Sullivan, D.J., Daly, D., O'Sullivan, O., Burdikova, Z., Vana, R., Beresford, T.P., Ross, R.P., Fitzgerald, G.F., McSweeney, P.L.H., et al. (2016). Thermus and the Pink Discoloration Defect in Cheese. mSystems *1*, e00023-16. https://doi.org/10.1128/mSystems.00023-16.
72. Salvetti, E., Campanaro, S., Campedelli, I., Fracchetti, F., Gobbi, A., Tornielli, G.B., Torriani, S., and Felis, G.E. (2016). Whole-Metagenome-Sequencing-Based Community Profiles of Vitis vinifera L. cv. Corvina Berries Withered in Two Post-harvest Conditions. Front. Microbiol. *7*, 937. https://doi.org/10.3389/fmicb.2016.00937.
73. Somerville, V., Lutz, S., Schmid, M., Frei, D., Moser, A., Irmler, S., Frey, J.E., and Ahrens, C.H. (2019). Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. BMC Microbiol. *19*, 143. https://doi.org/10.1186/s12866-019-1500-0.
74. Sternes, P.R., Lee, D., Kutyna, D.R., and Borneman, A.R. (2017). A combined meta-barcoding and shotgun metagenomic analysis of spontaneous wine fermentation. GigaScience *6*, 1–10.
75. Sulaiman, J., Gan, H.M., Yin, W.-F., and Chan, K.-G. (2014). Microbial succession and the functional potential during the fermentation of Chinese soy sauce brine. Front. Microbiol. *5*, 556. https://doi.org/10.3389/fmicb.2014.00556.

76. Walsh, A.M., Crispie, F., Kilcawley, K., O'Sullivan, O., O'Sullivan, M.G., Claesson, M.J., and Cotter, P.D. (2016). Microbial Succession and Flavor Production in the Fermented Dairy Beverage Kefir. mSystems *1*, e00052-16. https://doi.org/10.1128/mSystems.00052-16.

77. Walsh, A.M., Crispie, F., Daari, K., O'Sullivan, O., Martin, J.C., Arthur, C.T., Claesson, M.J., Scott, K.P., and Cotter, P.D. (2017). Strain-Level Metagenomic Analysis of the Fermented Dairy Beverage Nunu Highlights Potential Food Safety Risks. Appl. Environ. Microbiol. *83*, e01144-17. https://doi.org/10.1128/AEM.01144-17.

78. Xie, M., Wu, J., An, F., Yue, X., Tao, D., Wu, R., and Lee, Y. (2019). An integrated metagenomic/metaproteomic investigation of microbiota in dajiang-meju, a traditional fermented soybean product in Northeast China. Food Res. Int. *115*, 414–424. https://doi.org/10.1016/j.foodres.2018.10.076.

79. Yasir, M., Bibi, F., Hashem, A.M., and Azhar, E.I. (2020). Comparative metagenomics and characterization of antimicrobial resistance genes in pasteurized and homemade fermented Arabian laban. Food Res. Int. *137*, 109639. https://doi.org/10.1016/j.foodres.2020.109639.

80. Zhao, C.-C., and Eun, J.-B. (2020). Shotgun metagenomics approach reveals the bacterial community and metabolic pathways in commercial hongeo product, a traditional Korean fermented skate product. Food Res. Int. *131*, 109030. https://doi.org/10.1016/j.foodres.2020.109030.

81. Gänzle, M. (2022). The periodic table of fermented foods: limitations and opportunities. Appl. Microbiol. Biotechnol. *106*, 2815–2826. https://doi.org/10.1007/s00253-022-11909-y.

82. Tamang, J.P., Watanabe, K., and Holzapfel, W.H. (2016). Review: Diversity of Microorganisms in Global Fermented Foods and Beverages. Front. Microbiol. *7*, 377. https://doi.org/10.3389/fmicb.2016.00377.

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

84. Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L.J., Thompson, K.N., Zolfo, M., Manghi, P., Dubois, L., Huang, K.D., Thomas, A.M., et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn. Nat. Biotechnol. *41*, 1633–1644.

85. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132. https://doi.org/10.1186/s13059-016-0997-x.

86. Beghini, F., McIver, L.J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A.M., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife *10*, e65088. https://doi.org/10.7554/eLife.65088.

87. George, F., Daniel, C., Thomas, M., Singer, E., Guilbaud, A., Tessier, F.J., Revol-Junelles, A.-M., Borges, F., and Foligné, B. (2018). Occurrence and Dynamism of Lactic Acid Bacteria in Distinct Ecological Niches: A Multifaceted Functional Health Perspective. Front. Microbiol. *9*, 2899. https://doi.org/10.3389/fmicb.2018.02899.

88. Waśkiewicz, A., and Irzykowska, L. (2014). Flavobacterium spp. – Characteristics, Occurrence, and Toxicity. In Encyclopedia of Food Microbiology, Second Edition, C.A. Batt and M.L. Tortorello, eds. (Academic Press), pp. 938–942.

89. Lee, J.-Y., Kang, W., Kim, P.S., Lee, S.-Y., Shin, N.-R., Sung, H., Lee, J.-Y., Yun, J.-H., Jeong, Y.-S., Han, J.E., et al. (2020). Flaviflexus ciconiae sp. nov., isolated from the faeces of the oriental stork, Ciconia boyciana. Int. J. Syst. Evol. Microbiol. *70*, 5439–5444. https://doi.org/10.1099/ijsem.0.004435.

90. Stanborough, T., Fegan, N., Powell, S.M., Tamplin, M., and Chandry, P.S. (2017). Insight into the Genome of Brochothrix thermosphacta, a Problematic Meat Spoilage Bacterium. Appl. Environ. Microbiol. *83*, e02786-16. https://doi.org/10.1128/AEM.02786-16.

91. Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., and Hugenholtz, P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. *50*, D785–D794. https://doi.org/10.1093/nar/gkab776.

92. Handley, K.M., and Lloyd, J.R. (2013). Biogeochemical implications of the ubiquitous colonization of marine habitats and redox gradients by Marinobacter species. Front. Microbiol. *4*, 136. https://doi.org/10.3389/fmicb.2013.00136.

93. Blasche, S., Kim, Y., Mars, R.A.T., Machado, D., Maansson, M., Kafkia, E., Milanese, A., Zeller, G., Teusink, B., Nielsen, J., et al. (2021). Metabolic cooperation and spatiotemporal niche partitioning in a kefir microbial community. Nat. Microbiol. *6*, 196–208. https://doi.org/10.1038/s41564-020-00816-5.

94. Alraddadi, F.A.J., Ross, T., and Powell, S.M. (2023). Evaluation of the microbial communities in kefir grains and kefir over time. Int. Dairy J. *136*, 105490. https://doi.org/10.1016/j.idairyj.2022.105490.

95. Ammor, S., Dufour, E., Zagorec, M., Chaillou, S., and Chevallier, I. (2005). Characterization and selection of Lactobacillus sakei strains isolated from traditional dry sausage for their potential use as starter cultures. Food Microbiol. *22*, 529–538. https://doi.org/10.1016/j.fm.2004.11.016.

96. Leroy, S., Lebert, I., Chacornac, J.-P., Chavant, P., Bernardi, T., and Talon, R. (2009). Genetic diversity and biofilm formation of Staphylococcus equorum isolated from naturally fermented sausages and their manufacturing environment. Int. J. Food Microbiol. *134*, 46–51. https://doi.org/10.1016/j.ijfoodmicro.2008.12.012.

97. Chen, Y., Yu, L., Qiao, N., Xiao, Y., Tian, F., Zhao, J., Zhang, H., Chen, W., and Zhai, Q. (2020). Latilactobacillus curvatus: A Candidate Probiotic with Excellent Fermentation Properties and Health Benefits. Foods *9*, 1366. https://doi.org/10.3390/foods9101366.

98. Yu, L., Chen, Y., Duan, H., Qiao, N., Wang, G., Zhao, J., Zhai, Q., Tian, F., and Chen, W. (2022). Latilactobacillus sakei: a candidate probiotic with a key role in food fermentations and health promotion. Crit. Rev. Food Sci. Nutr. *64*, 978–995.

99. Wang, G.-Y., Li, M., Ma, F., Wang, H.-H., Xu, X.-L., and Zhou, G.-H. (2017). Physicochemical properties of Pseudomonas fragi isolates response to modified atmosphere packaging. FEMS Microbiol. Lett. *364*, fnx106. https://doi.org/10.1093/femsle/fnx106.

100. De Filippis, F., La Storia, A., Villani, F., and Ercolini, D. (2019). Strain-Level Diversity Analysis of Pseudomonas fragi after In Situ Pangenome Reconstruction Shows Distinctive Spoilage-Associated Metabolic Traits Clearly Selected by Different Storage Conditions. Appl. Environ. Microbiol. *85*, e02212-18. https://doi.org/10.1128/AEM.02212-18.

101. Sarkar, P.K., Hasenack, B., and Nout, M.J.R. (2002). Diversity and functionality of Bacillus and related genera isolated from spontaneously fermented soybeans (Indian Kinema) and locust beans (African Soumbala). Int. J. Food Microbiol. *77*, 175–186. https://doi.org/10.1016/s0168-1605(02)00124-1.

102. Li, Z., Zheng, M., Zheng, J., and Gänzle, M.G. (2023). Bacillus species in food fermentations: an underappreciated group of organisms for safe use in food fermentations. Curr. Opin. Food Sci. *50*, 101007. https://doi.org/10.1016/j.cofs.2023.101007.

103. Yassour, M., Jason, E., Hogstrom, L.J., Arthur, T.D., Tripathi, S., Siljander, H., Selvenius, J., Oikarinen, S., Hyöty, H., Virtanen, S.M., et al. (2018). Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. Cell Host Microbe *24*, 146–154.e4. https://doi.org/10.1016/j.chom.2018.06.007.

104. Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., et al. (2015). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe *17*, 690–703. https://doi.org/10.1016/j.chom.2015.04.004.

105. Shao, Y., Forster, S.C., Tsaliki, E., Vervier, K., Strang, A., Simpson, N., Kumar, N., Stares, M.D., Rodger, A., Brocklehurst, P., et al. (2019). Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. Nature *574*, 117–121. https://doi.org/10.1038/s41586-019-1560-1.

106. Baek, M.-G., Kim, K.W., and Yi, H. (2023). Subspecies-level genome comparison of Lactobacillus delbrueckii. Sci. Rep. *13*, 3171. https://doi.org/10.1038/s41598-023-29404-3.

107. Partridge, S.R., Kwong, S.M., Firth, N., and Jensen, S.O. (2018). Mobile Genetic Elements Associated with Antimicrobial Resistance. Clin. Microbiol. Rev. *31*, e00088-17. https://doi.org/10.1128/CMR.00088-17.

108. Cao, Z., Cui, L., Liu, Q., Liu, F., Zhao, Y., Guo, K., Hu, T., Zhang, F., Sheng, X., Wang, X., et al. (2022). Phenotypic and Genotypic Characterization of Multidrug-Resistant Enterobacter hormaechei Carrying qnrS Gene Isolated from Chicken Feed in China. Microbiol. Spectr. *10*, e0251821. https://doi.org/10.1128/spectrum.02518-21.

109. Kamathewatta, K., Bushell, R., Rafa, F., Browning, G., Billman-Jacobe, H., and Marenda, M. (2020). Colonization of a hand washing sink in a veterinary hospital by an Enterobacter hormaechei strain carrying multiple resistances to high importance antimicrobials. Antimicrob. Resist. Infect. Control *9*, 163. https://doi.org/10.1186/s13756-020-00828-0.

110. Nandi, S.P., Sultana, M., and Hossain, M.A. (2013). Prevalence and characterization of multidrug-resistant zoonotic Enterobacter spp. in poultry of Bangladesh. Foodborne Pathog. Dis. *10*, 420–427. https://doi.org/10.1089/fpd.2012.1388.

111. Hoffmann, H., and Roggenkamp, A. (2003). Population genetics of the nomenspecies Enterobacter cloacae. Appl. Environ. Microbiol. *69*, 5306–5318. https://doi.org/10.1128/AEM.69.9.5306-5318.2003.

112. Thomas, A.M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., Beghini, F., Manara, S., Karcher, N., Pozzi, C., et al. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat. Med. *25*, 667–678. https://doi.org/10.1038/s41591-019-0405-7.

113. Boleij, A., van Gelder, M.M.H.J., Swinkels, D.W., and Tjalsma, H. (2011). Clinical Importance of Streptococcus gallolyticus infection among colorectal cancer patients: systematic review and meta-analysis. Clin. Infect. Dis. *53*, 870–878. https://doi.org/10.1093/cid/cir609.

114. Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X., Jie, Z., et al. (2015). Gut microbiome development along the colorectal adenoma–carcinoma sequence. Nat. Commun. *6*, 6528. https://doi.org/10.1038/ncomms7528.

115. Maixner, F., Sarhan, M.S., Huang, K.D., Tett, A., Schoenafinger, A., Zingale, S., Blanco-Míguez, A., Manghi, P., Cemper-Kiesslich, J., Rosendahl, W., et al. (2021). Hallstatt miners consumed blue cheese and beer during the Iron Age and retained a non-Westernized gut microbiome until the Baroque period. Curr. Biol. *31*, 5149–5162.e6. https://doi.org/10.1016/j.cub.2021.09.031.

116. Gallone, B., Steensels, J., Prahl, T., Soriaga, L., Saels, V., Herrera-Malaver, B., Merlevede, A., Roncoroni, M., Voordeckers, K., Miraglia, L., et al. (2016). Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. Cell *166*, 1397–1410.e16. https://doi.org/10.1016/j.cell.2016.08.020.

117. Al-Qaysi, S.A.S., Al-Haideri, H., Thabit, Z.A., Al-Kubaisy, W.H.A.A.-R., and Ibrahim, J.A.A.-R. (2017). Production, Characterization, and Antimicrobial Activity of Mycocin Produced by Debaryomyces hansenii DSMZ70238. Int. J. Microbiol. *2017*, 2605382. https://doi.org/10.1155/2017/2605382.

118. Ramos-Moreno, L., Ruiz-Pérez, F., Rodríguez-Castro, E., and Ramos, J. (2021). Debaryomyces hansenii Is a Real Tool to Improve a Diversity of Characteristics in Sausages and Dry-Meat Products. Microorganisms *9*, 1512. https://doi.org/10.3390/microorganisms9071512.

119. Parapouli, M., Vasileiadis, A., Afendra, A.-S., and Hatziloukas, E. (2020). Saccharomyces cerevisiae and its industrial applications. AIMS Microbiol. *6*, 1–31. https://doi.org/10.3934/microbiol.2020001.

120. Duan, S.-F., Han, P.-J., Wang, Q.-M., Liu, W.-Q., Shi, J.-Y., Li, K., Zhang, X.-L., and Bai, F.-Y. (2018). The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. Nat. Commun. *9*, 2690. https://doi.org/10.1038/s41467-018-05106-7.

121. Pontes, A., Hutzler, M., Brito, P.H., and Sampaio, J.P. (2020). Revisiting the Taxonomic Synonyms and Populations of Saccharomyces cerevisiae-Phylogeny, Phenotypes, Ecology and Domestication. Microorganisms *8*, 903. https://doi.org/10.3390/microorganisms8060903.

122. Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., Armanini, F., Truong, D.T., Manara, S., Zolfo, M., et al. (2018). Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host Microbe *24*, 133–145.e5. https://doi.org/10.1016/j.chom.2018.06.005.

123. Brauer, A., and Bengtsson, M.M. (2022). DNA extraction bias is more pronounced for microbial eukaryotes than for prokaryotes. MicrobiologyOpen *11*, e1323. https://doi.org/10.1002/mbo3.1323.

124. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

125. Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol. Biol. Evol. *38*, 4647–4654. https://doi.org/10.1093/molbev/msab199.

126. Asnicar, F., Weingart, G., Tickle, T.L., Huttenhower, C., and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ *3*, e1029. https://doi.org/10.7717/peerj.1029.

127. Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics *31*, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033.

128. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ *7*, e7359. https://doi.org/10.7717/peerj.7359.

129. Asnicar, F., Thomas, A.M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., Zhu, Q., Bolzan, M., Cumbo, F., May, U., et al. (2020). Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nat. Commun. *11*, 2500. https://doi.org/10.1038/s41467-020-16366-7.

130. R Core Team (2020). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

131. The Scikit-Bio Development Team (2020). scikit-bio: A Bioinformatics Library for Data Scientists, Students, and Developers. https://scikit.bio/.

132. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

133. Seabold, S., and Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In Proceedings of the 9th Python in Science Conference (SciPy), pp. 92–96. https://doi.org/10.25080/Majora-92bf1922-011.

134. Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., et al. (2022). vegan: Community Ecology Package. R package version 2.5–7.

135. Barcenilla, C., Cobo-Díaz, J.F., De Filippis, F., Valentino, V., Cabrera Rubio, R., O'Neil, D., Mahler de Sanchez, L., Armanini, F., Carlino, N., Blanco-Míguez, A., et al. (2024). Improved sampling and DNA extraction procedures for microbiome analysis in food-processing environments. Nat. Protoc. *19*, 1291–1310. https://doi.org/10.1038/s41596-023-00949-x.

136. Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. *35*, 725–731. https://doi.org/10.1038/nbt.3893.

137. Legendre, P., and Anderson, M.J. (1999). Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. Ecol. Monogr. *69*, 1–24. https://doi.org/10.1890/0012-9615(1999)069[0001:DBRATM]2.0.CO;2.

138. Blanchet, F.G., Legendre, P., and Borcard, D. (2008). Forward selection of explanatory variables. Ecology *89*, 2623–2632. https://doi.org/10.1890/07-0986.1.

139. Pasolli, E., Truong, D.T., Malik, F., Waldron, L., and Segata, N. (2016). Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLoS Comput. Biol. *12*, e1004977. https://doi.org/10.1371/journal.pcbi.1004977.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Food samples	This paper	N/A
Critical commercial assays		
PowerBead Pro Tubes	Qiagen, Germany	Catalog No. 19301
DNeasy PowerSoil Pro kit	Qiagen, Germany	Catalog No. 47014
Nextera DNA Flex Library Prep	Illumina, California, USA	Catalog No. 20018705
Deposited data		
Raw sequencing data (food metagenomes)	This paper	See Table S1
MAGs (from food metagenomes)	This paper	https://zenodo.org/doi/10.5281/zenodo.10891046
Software and algorithms		
BowTie2 (version 2.2.9)	Langmead and Salzberg[124]	https://github.com/BenLangmead/bowtie2
BUSCO (version 5.6.1)	Manni et al.[125]	https://gitlab.com/ezlab/busco
CheckM (version 1.1.3)	Parks et al.[83]	https://github.com/Ecogenomics/CheckM
ChocoPhlAn (version 2022.04)	Blanco-Míguez et al.[84]	https://github.com/biobakery/humann
ComplexHeatmap (version 2.18)	N/A	https://github.com/jokergoo/ComplexHeatmap
corrplot (version 0.90)	N/A	https://github.com/taiyun/corrplot
dplyr (version 1.1.4)	N/A	https://github.com/tidyverse/dplyr
ggplot2 (version 3.4.2)	N/A	https://github.com/tidyverse/ggplot2
GraphlAn (version 1.1.3)	Asnicar et al.[126]	https://bitbucket.org/nsegata/graphlan/
HUMAnN (version 3)	Beghini et al.[86]	https://github.com/biobakery/humann
Mash (version 2)	Ondov et al.[85]	https://github.com/marbl/Mash
MEGAHIT (version 1.1.1)	Li et al.[127]	https://github.com/voutcn/megahit
MetaBAT (version 2.12.1)	Kang et al.[128]	https://bitbucket.org/berkeleylab/metabat
MetaPhlAn (version 4)	Blanco-Míguez et al.[84]	https://github.com/biobakery/MetaPhlAn
numpy (version 1.26.0)	N/A	https://github.com/numpy/numpy
pandas (version 2.1)	N/A	https://github.com/pandas-dev/pandas
PyPhlAn	N/A	https://github.com/SegataLab/pyphlan
PhyloPhlAn (version 3)	Asnicar et al.[129]	https://github.com/biobakery/phylophlan
R	R Core Team[130]	R: The R Project for Statistical Computing (r-project.org)
scikit-bio (version 0.5.6)	The Scikit-Bio Development Team[131]	https://github.com/scikit-bio/scikit-bio
scipy (version 1.5.3)	Virtanen et al.[132]	https://github.com/scipy/scipy
stats (version 4.0.3)	R Core Team[130]	NA
statsmodel (version 0.13.1)	Seabold and Perktold[133]	https://github.com/statsmodels/statsmodels
StrainPhlAn (version 4)	Blanco-Míguez et al.[84]	https://github.com/biobakery/MetaPhlAn
Trim Galore (version 0.6.6)	N/A	https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
vegan (version 2.6-4)	Oksanen et al.[134]	https://github.com/vegandevs/vegan
Other		
curatedFoodMetagenomicData (cFMD)	This paper	https://github.com/SegataLab/cFMD
curatedMetagenomicData (cMD)	Pasolli et al.[27]	https://waldronlab.github.io/curatedMetagenomicData/
NCBI GenBank database	N/A	https://www.ncbi.nlm.nih.gov/genbank/

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to and will be fulfilled by the lead contact, Nicola Segata (nicola.segata@unitn.it).

Materials availability

All materials used in this study are commercially available, as specified in the key resources table.

Data and code availability

- cFMD is freely available in the GitHub repository (https://github.com/SegataLab/cFMD) and metadata are contextually available within this paper as Table S1.
- Raw sequencing data for the full set of metagenomes analysed are publicly available and can be retrieved using the accession numbers reported in Table S1. The food MAGs are publicly available at https://zenodo.org/doi/10.5281/zenodo.10891046.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Overview of the resource

cFMD is a public database in the form of a GitHub repository (https://github.com/SegataLab/cFMD) containing curated metadata in addition to taxonomic and functional profiles for thousands of food (shotgun) metagenomes. The first version of cFMD consists of a total of 2,533 metagenomes associated with 59 datasets: 45 datasets and 583 samples are coming from publicly available studies, and the remaining 14 datasets and 1,950 samples are produced inside the MASTER EU Consortium and made available with this paper.

Survey of public datasets associated with food metagenomes

We conducted a thorough literature survey to screen publicly available studies associated with food metagenomes. We searched for papers and publications through multiple web services (i.e., Pubmed, ResearchGate, BiorXiv, and editors archives), in addition to collecting raw data directly from sequence databases (i.e., NCBI, MG-RAST, GSA). We preliminarily identified over 100 studies for shotgun metagenomics, which were filtered out by keeping only projects with whole-genome sequencing (WGS), full availability of raw data and metadata, sequencing through Illumina technology, and satisfying a first check on the quality of the sequencing. This resulted in 45 datasets for a total of 583 samples (average sequencing = 3.0 ± 5.8 Gb/sample and 21.6M ± 43.3M reads/sample).

Collection of new food metagenomes

We additionally considered 1,950 metagenomes spanning 14 datasets that were produced thanks to the MASTER EU Consortium and that are now, contextually with this paper, publicly available. Here, we kept only food samples for the purposes of cFMD, while additional control (n = 141), environmental (n = 30), and factory (n = 752) samples were analysed elsewhere.[135]

Integration of cFMD with human metagenomes from cMD

We also conducted a large-scale analysis aiming at finding overlaps between food and human microbial communities. The set of food metagenomes previously described was complemented by a large amount of human data from publicly available sources. We considered the already collected metagenomes and standardised metadata information available in cMD.[27] We included all human metagenomes from oral and stool sources available as of May 2022, which resulted in a total of 19,833 samples from 87 cohorts. These samples were categorised in terms of body site (i.e., oral and stool) and host lifestyle (i.e., W: westernised and NW: non-westernised). We therefore defined four main groups as follows: stool_W (n = 17,884), stool_NW (n = 1,092), oral_W (n = 694), and oral_NW (n = 163). These samples were also subdivided in terms of age category as defined in Pasolli et al.[27]: newborn (age <1 years old, n = 2,892), child (1<= age <12, N = 1,175), school age (12<= age <19, N = 720), adult (19<= age <= 65, N = 13,334), and senior (age >65, N = 1,712).

METHOD DETAILS

Metadata curation and standardization

The 2,533 food metagenomes were acquired globally (i.e., 50 countries and 5 continents). Most represented continents were Europe (n = 1,995), Asia (n = 183), and North America (n = 183). For each sample, we collected a total of 27 metadata fields with syntactic rules summarised in Table S1: 15 mandatory fields (e.g., dataset_name, sample_id, macrocategory, category, type, and sequencing_platform) and 12 optional ones. Such 27 fields can be further categorised as follows: sample-related (n = 3 fields; i.e., dataset_name, sample_id, and country), food-related (n = 6 fields; i.e., macrocategory, category, type, subtype, commercial

name, and fermented/non-fermented), and technical (n = 18 fields; i.e., sample_accession, run_accession, experiment_accession, study_accession, project_accession, database_origin, library_layout, sequencing_platform, DNA_extraction_kit, collection_date, n_of_bases, n_of_reads, min_read_len, median_read_len, mean_read_len, max_read_len, filtered, and curator) characteristics. This defined metadata was based on previous work in the field,[81] and further expanded by food microbiologists as part of the current work. Such information was retrieved from the original papers/databases or collected within the MASTER EU Consortium. The unique key for querying the database is represented by the dataset_name and sample_id.

Samples were classified according to their composition and production using three hierarchical levels of detail: 15 categories (i.e., alcoholic beverages, dairy, (non-alcoholic) fermented beverages, fermented fish, fermented fruits and vegetables, fermented grains, fermented legumes, fermented meat, fermented seeds, fermented tubers and roots, (non-fermented) fish, (non-fermented) fruits and vegetables, (non-fermented) meat, other, and probiotics - intended as dietary supplements) gathering 107 types and 358 subtypes. In the ''fermented/non-fermented'' field, samples were classified in two groups: fermented (F) and non-fermented (NF).

DNA isolation, library preparation, and sequencing of food samples

The metagenomes generated within the MASTER EU Consortium underwent a recently developed protocol (similar to what described in Barcenilla et al.[135]) aiming at maximising the collected microbial DNA from the generally low biomass samples. Commons steps comprised: i) cell lysis occurring through a combination of both mechanical (bead beating in Qiagen's PowerBead Pro Tubes) and chemical (Qiagen DNeasy PowerSoil Pro kit) methods; ii) library preparation based on the Illumina Nextera DNA Flex Library Prep following the manufacture protocol; and iii) libraries multiplexed using dual indexing and sequenced for 300 bp paired-end reads on the Illumina NovaSeq 6000 Sequencing System.

Pre-processing of raw-reads

Raw reads of the generated food samples were pre-processed through a validated pipeline (available at https://github.com/SegataLab/preprocessing) based on these subsequent main steps: i) discarding of low-quality (quality<20), short (L<75bp) and with too many ambiguous nucleotides (N>2) reads using Trim Galore v0.6.6 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/); ii) removal of human (hg19 human genome release), bacteriophage phiX174 DNA (Illumina spike-in), and other hosts (where applicable) contamination by mapping reads against their reference genomes through BowTie2 v2.2.9 (with parameter –sensitive-local)[124] iii) sorting and splitting of the reads passing the filtering into forward, reverse, and unpaired files.

Downstream analyses were performed by discarding the 21 samples having less than 0.1 quality-controlled Gbases (marked as filtered=yes in Table S1) resulting in a total of 2,512 food metagenomes.

Extraction of prokaryotic and eukaryotic MAGs

MAGs were generated by applying *de novo* metagenomic assembly to each metagenome independently. This was done by considering the pipeline proposed and validated for prokaryotes in Pasolli et al.[28]: 1) metagenomic assembly with MEGAHIT v1.1.1[127]; 2) removal of contigs shorter than 1000 bp; 3) alignment of the remaining contigs against original raw data with Bowtie2 v2.2.9[124] to calculate coverage information; 4) binning of the contigs through MetaBAT v2.12.1[128]; 5) prokaryotic quality control of the resulting putative genomes with CheckM[83] v1.1.3 and successive filtering of high-quality (completeness>90% and contamination<5%; HQ) and medium-quality (completeness$\geq$50% and contamination<5%; MQ) MAGs according to the standard proposed in Bowers et al.[136]; 6) on the low quality (LQ) MAGs resulting from 5), eukaryotic quality control with BUSCO v5.6.1[125]; the same thresholds adopted in 5) were considered to define MQ and HQ eukaryotic MAGs. This resulted in a total of 5,136 HQ and 4,976 MQ prokaryotic MAGs and 392 HQ and 395 MQ eukaryotic MAGs.

Expansion of MetaRefSGB and ChocoPhlAn with the extracted MAGs

The MAGs that satisfied the quality check were used to expand the collection of MAGs and isolate genomes available in MetaRefSGB. In this resource, ANIs among quality-controlled genomes were computed through Mash[85] and genomes having a genetic identity greater than 95% were clustered into the same SGB. Redundancy was avoided by filtering out genomes with a genetic distance < 0.01%. An SGB was defined as known SGB (kSGB) if it contained at least one genome from isolate sources, otherwise it was named as unknown SGB (uSGB). We expanded this resource (labelled as vMar22) to a total of 1.17M MAGs and 80K distinct SGBs. Our 10,112 prokaryotic food MAGs contributed to 535 kSGBs and 211 uSGBs, with additional 290 ufSGBs that were defined as food-specific thanks to our set of extracted MAGs. Similarly, our 787 eukaryotic food MAGs contributed to 78 kSGBs and 30 uSGBs.

Taxonomic labelling of each prokaryotic SGB was performed by considering the strategy adopted in Pasolli et al.[28]: 1) a kSGB was labelled with the species label associated with the reference genome(s) present in the SGB; majority rule was applied in the case of multiple reference genomes with different taxonomies; 2) assignment at higher taxonomic level was provided for uSGBs, which do not include any reference genomes for definition. We defined genus-level genome bins (GGBs) by clustering genomes at 85% ANI, and the same majority rule adopted in 1) was considered to assign genus-level taxonomy. The same procedure was also applied for family level genomes bins (FGBs), in which genomes were clustered at 70% ANI. The same procedure was also extended to eukaryotic SGBs, which were taxonomically labelled based on 17,438 publicly available genomes.

The expanded MetaRefSGB resource was used to build the database of SGB-specific marker genes through the ChocoPhlAn pipeline[84] (Figure 1E). Only SGBs having at least either one reference genome (i.e. a genome sequenced from an isolate) or 5 MAGs in MetaRefSGB were kept in the ChocoPhlAn database. This expanded the profilable prokaryotes from 26,970 SGBs (21,978 kSGBs and 4,992 uSGBs; published vJan21[84]) to 29,480 SGBs (22,020 kSGBs and 7,460 uSGBs). Similarly, the catalogue of Eukaryotes was extended from 122 species (17 genera) to 489 species (136 genera). Such an improved SGB-centric unique marker database was used for species-level taxonomic profiling and strain-level analysis as described in the sections ''generation and analysis of taxonomic and functional profiles'' and ''strain-level characterization of common prevalent SGBs through StrainPhlAn'', respectively.

Phylogenetic analysis for building the bacterial tree of life

We retrieved a large set of genomes to cover the diversity of human and food microbiomes. We identified the 1,036 prokaryotic SGBs containing at least one MAG from food; additionally, we considered 3,962 prokaryotic SGBs prevalent in human microbiomes (i.e., with >=3 human MAGs) and without any MAGs from food (Table S7). This resulted in a total of 4,998 prokaryotic SGBs, and we selected one representative MAG per SGB. For kSGBs, this was chosen by selecting randomly one genome from those coming from isolate sequencing; for uSGBs it was chosen empirically based on CheckM[83] estimates as the MAG maximising the value (completeness-3*contamination).

The phylogenetic tree of life (Figure 2) was built through PhyloPhlAn v3[129] and by considering the 400 universal markers available in PhyloPhlAn. Parameters were set as follows: ''–diversity high –fast –min_num_markers 50''. Ten SGBs were discarded from the tree since their representatives contained less than 50 universal marker genes. The resulting tree was rooted on the Archaea and visualised with GraPhlAn.[126]

Generation of taxonomic and functional profiles

Read mapping-based profiles were generated through the bioBakery suite.[86] More specifically, SGB-level taxonomic profiles were obtained through MetaPhlAn v4.0.2 (qstat = 0.2)[84] and by considering the updated marker database (v202204) available in ChocoPhlAn and obtained as described in ''Expansion of MetaRefSGB and ChocoPhlAn with the extracted MAGs''. We obtained a non-empty taxonomic profile for 2,482 food metagenomes, which was considered as the final set of samples for downstream analyses. We found 3,622 SGBs present in at least one food metagenome, with an average of 25 ± 30 SGBs per sample. Also the taxonomic profiles of the 19,833 human metagenomes were generated using the same pipeline. The final analysis was conducted on the set of 19,786 human metagenomes by excluding the 47 ones having an empty taxonomic profile. Functional profiles were obtained through HUMAnN v3[86] using the UniRef90 pangenomes and by considering the full set of species detected by MetaPhlAn v3[86] (–metaphlan-options "-t rel_ab –index v30_CHOCOPhlAn_201901''). We generated UniRef90 gene family abundance data as well as metabolic pathway abundance and coverage.

QUANTIFICATION AND STATISTICAL ANALYSIS

Analysis of taxonomic profiles

For the analysis reported in Figure 4A, intra-sample diversity (i.e., alpha diversity) was computed in terms of estimated richness and Shannon diversity index. Due to the significant correlation between alpha diversity and sequencing depth (for estimated richness: $p < 2.2e-16$, intercept: (15.6 ± 0.8), slope: (1.56 ± 0.09)e-09; for Shannon diversity index: $p < 1.28e-06$, intercept: (1.09 ± 0.02), slope: (1.1 ± 0.2)e-11), their statistical significance across food categories was evaluated by, first, correcting their values by sequencing depth through a linear model with the parameters above and, second, by considering Wilcoxon rank sum test and false discovery rate (Benjamini-Hochberg FDR) correction (R package stats v4.0.3,[130] Figures S3G and S3H).

Inter-sample diversity (i.e., beta diversity, Figure 4B) was assessed in terms of Bray-Curtis distances using the script available in MetaPhlAn v4 release and additionally performing ordination analysis by applying t-distributed stochastic neighbour embedding (t-SNE).[84] The partial distance-based redundancy analysis[137] was run on metadata variables to explain Bray-Curtis distances among samples using the R vegan package v2.6-4.[134] The analysis was run on the 2,333 metagenomes having a non-empty taxonomic profile and non-missing values in the metadata fields. We show in Figure 4C only the eight variables (i.e., category, type, and subtype from food ontology, continent, and country as sample geographic description and dataset_name, sequencing_platform, and DNA_extraction_kit to take into account technical and batch effects) significant (FDR corrected p value <0.001) in univariate analysis. Cumulative analysis was additionally assessed by performing a permutation test (n = 1000) on the constrained ordination[138] using the ordiR2step function and a stepwise forward approach (R vegan package v2.6-4).[134] This cumulative setting discarded three variables (i.e., category, continent, and DNA_extraction_kit). Statistical significance across food categories was performed through analysis of similarities (ANOSIM; python package scikit-bio v0.5.6[131]) and permutational multivariate analysis of variance (PERMANOVA[131]) corrected by FDR (Python package statsmodel v0.13.1[133]). The top-25 prevalent SGBs were identified by considering the 90th percentile in terms of rel. ab. (Figure 4D). Separability among food categories was also assessed through a machine learning-based classification approach. For each category, we considered a binary classification setting aiming at discriminating that specific category from the rest using species-level taxonomic profiles. We considered random forest (RF) as back-end classifier in a repeated cross-validation implementation (5 folds and 5 repeats) as originally proposed in Pasolli et al.[139] Classification accuracies were assessed in

terms of AUC. The same machine-learning approach was extended to discrimination among dairy samples; we considered only food types with >=5 non-empty taxonomic profiles (n = 1,606 samples from 8 different dairy types). Similarly, we performed subtype classification by restricting the analysis to cheese samples; we considered only subtypes with at least 5 non-empty taxonomic profiles and associated with commercially available products (n = 293 samples from 23 cheese subtypes).

We identified SGBs differentially prevalent between food categories by applying Fisher's exact test (scipy python package version 1.5.3[132]) on contingency tables generated by presence/absence profiles (Figures 4E, 4F, and S4A–S4C). We also defined the ''SGB enrichment score'' as a metric to summarise the number of SGBs differentially prevalent between two food categories. The ''SGB enrichment score'' between two food categories *c1* and *c2* was computed as follows: i) compute the number of SGBs enriched in *c1* (with respect to *c2*) based on the Fisher's exact test; ii) similarly, compute the number of SGBs enriched in *c2* (with respect to *c1*); iii) compute the score as the difference between the two numbers calculated in points i) and ii). The score is thus a positive number when more SGBs are enriched in category *c1* with respect to *c2* (and vice versa for a negative value of the score).

Integrative analysis of food and human metagenomes

We defined the food-prevalent SGBs as those found with rel. ab. $\geq$ 0.1% in $\geq$4 food samples from taxonomic profiles. In this way, we identified 816 food-prevalent species. Although the choice of 0.1% as threshold was empirical, the number of prevalent SGBs varied only slightly when considering higher threshold values (762 for 0.5% and 711 for 1%). Among these 816 food-prevalent SGBs, 409 of them were also detected in at least one human metagenome. We analysed the distribution of these 409 SGBs across human samples according to lifestyle, body site, and age category in terms of: absolute richness (i.e., count of food-prevalent SGBs, Figure S5C); richness ratio (normalised by the total richness of each sample, Figure 5B); and cumulative rel. ab (Figure 5C). Statistical difference among human groups was calculated using Wilcoxon rank sum test and false discovery rate (Benjamini-Hochberg FDR) correction (R package stats v4.0.3[130]). We also defined as human-prevalent the SGBs found with rel. ab. $\geq$ 0.1% in $\geq$1% of the samples in at least one of the four human sub-groups (i.e., stool_W, stool_NW, oral_W, and oral_NW). This resulted in a total of 1749 human-prevalent SGBs. Finally, we restricted the analysis to the 43 SGBs prevalent in both human and food metagenomes.

Strain-level characterization of common prevalent SGBs through StrainPhlAn

We generated strain-level profiles on the 43 SGBs commonly prevalent between food and human sources through StrainPhlAn 4.[84,86] StrainPhlAn 4 was run independently on each SGB by considering the set of samples in which MetaPhlAn v4 detected the SGB under investigation with a rel. ab. > 0. All available reference genomes were also included. The parameters were set as follows: –marker_in_n_samples 66; –sample_with_n_markers 66; –sample_with_n_markers_after_filt 50; –phylophlan_mode accurate; –mutation_rates; and –debug. We finally kept the 16 SGBs (Figures 6 and S6) for which StrainPhlAn kept more than 15 food samples with a length in terms of multiple sequence alignment greater than 3kb. For each resulting phylogenetic tree and leaf pair, pairwise distances were calculated with the PyPhlAn package (https://github.com/SegataLab/pyphlan) and normalised by the total branch length.

The strain-level analysis for *S. cerevisiae* (Figure 7C) was extended by also considering 157 isolate reference genomes from a recent study of mostly Liquid State Fermentation (LSF) yeast.[115] The StrainPhlAn parameters were set as follows: –marker_in_n_samples 30; –sample_with_n_markers 100; –sample_with_n_markers_after_filt 50; –abs_n_markers_thres; –trim_sequence 25; –breadth_thres 50; –phylophlan_mode accurate; –mutation_rates; and –debug. The tree was rooted using Wild and Solid State Fermentation (SSF). To test for the robustness of our phylogenetic inference, we bootstrapped the alignment (Figure 7C, node bootstrap results shown in Figure S7A) and repeated the analysis using a more stringent selection of sites (Figure S7B).

CellPress
OPEN ACCESS

Supplemental figures

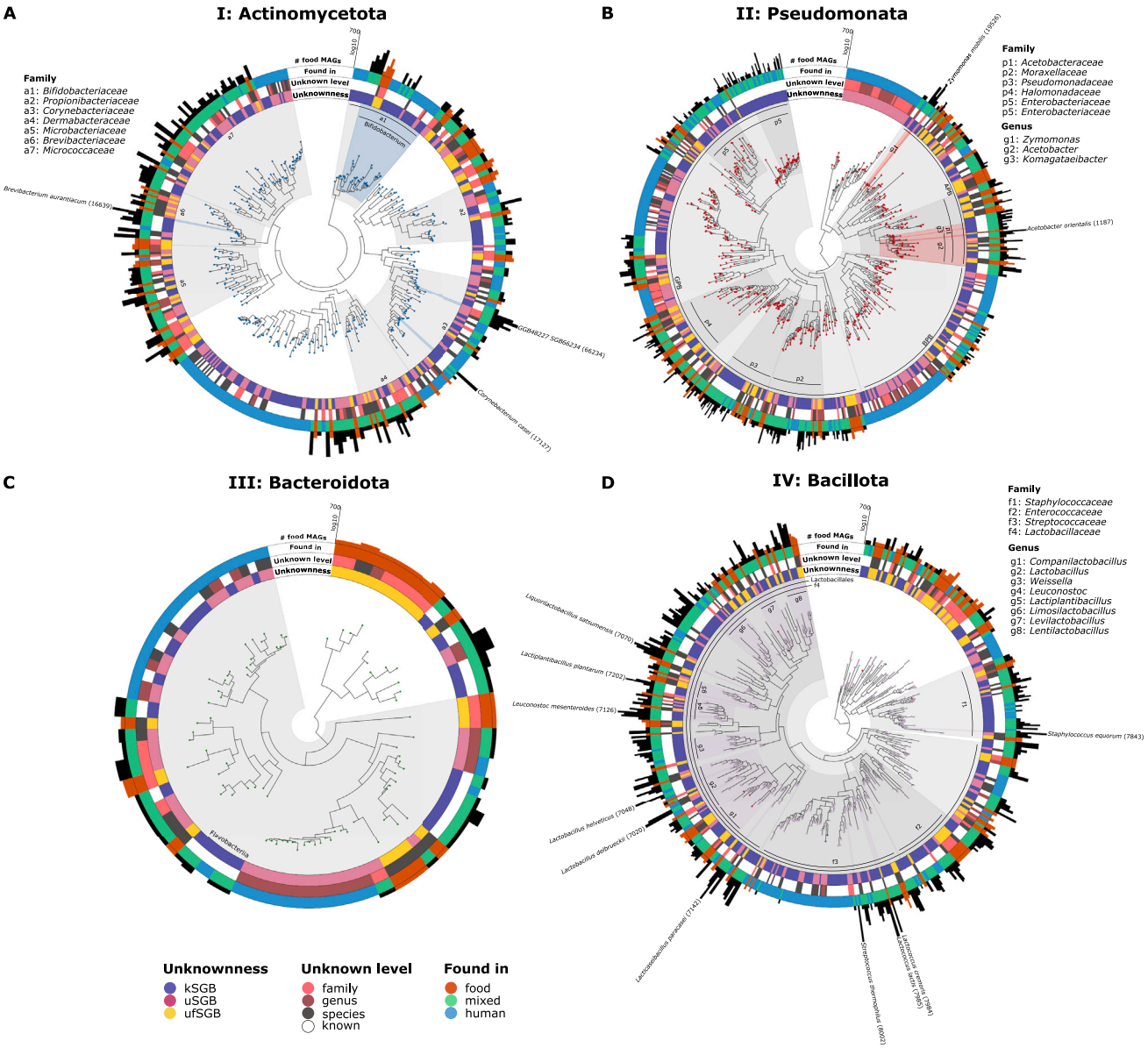
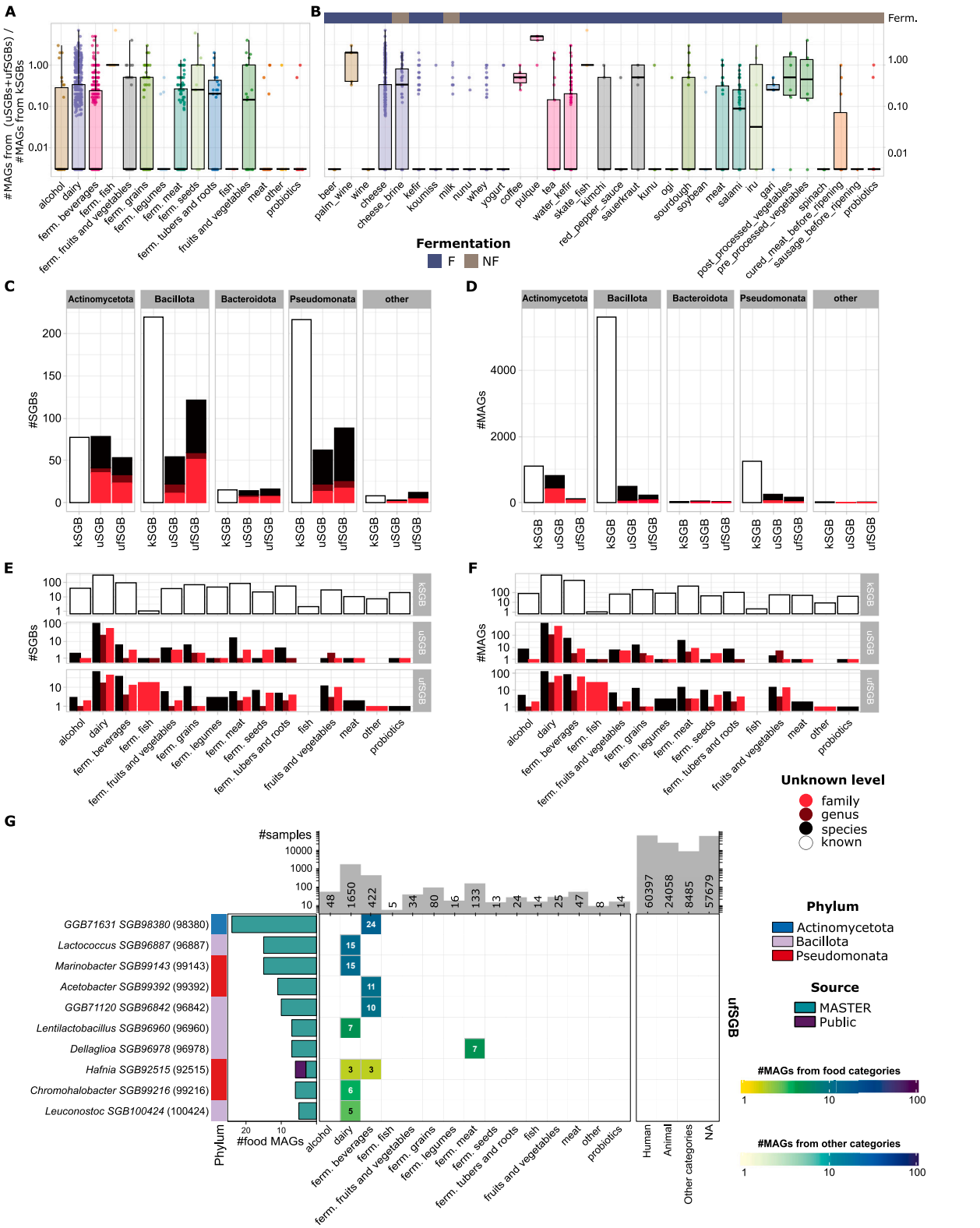


Figure S1. Phylogenetic trees for specific clades prevalent in food metagenomes, related to Figure 2

The phylogenetic tree shown in Figure 2 is detailed here for clades of relevance in food: (A) I: Actinomycetota; (B) II: Pseudomonadota; (C) III: Bacteroidota; (D) IV: Bacillota. Leaves are colored according to phylum as in Figure 2A. SGB labels radiating from the tree represent important species and are highlighted in the phylogenies by stars.



(legend on next page)

Figure S2. The reconstructed MAGs enlarge the set of available SGBs, related to Figures 2 and 3

(A–F) Ratio between the number of MAGs belonging to uSGBs + ufSGBs over MAGs assigned to kSGBs per sample according to (A) food category and (B) food type, with the indication of fermented (F) and non-fermented (NF) types. Only food types with $\geq$5 samples are shown. kSGB is a cluster of genomes including at least one reference genome; uSGB does not include any reference genomes but was previously reconstructed from non-food sources; ufSGB does not include reference genomes and was only identified in food. Number of (C) SGBs and (D) MAGs stratified by phylum, unknownness (i.e., kSGB, uSGB, and ufSGB), and unknown level. Similarly, the number of (E) SGBs and (F) MAGs is reported by stratifying for food category, unknownness, and unknown level.

(G) The ten ufSGBs with the highest number of MAGs. The number of food MAGs is reported in total and for each of the 15 categories defined in this study. For definition, no other MAGs retrieved from other sources (i.e., human, animal, and other categories) are available in the MetaRefSGB repository.

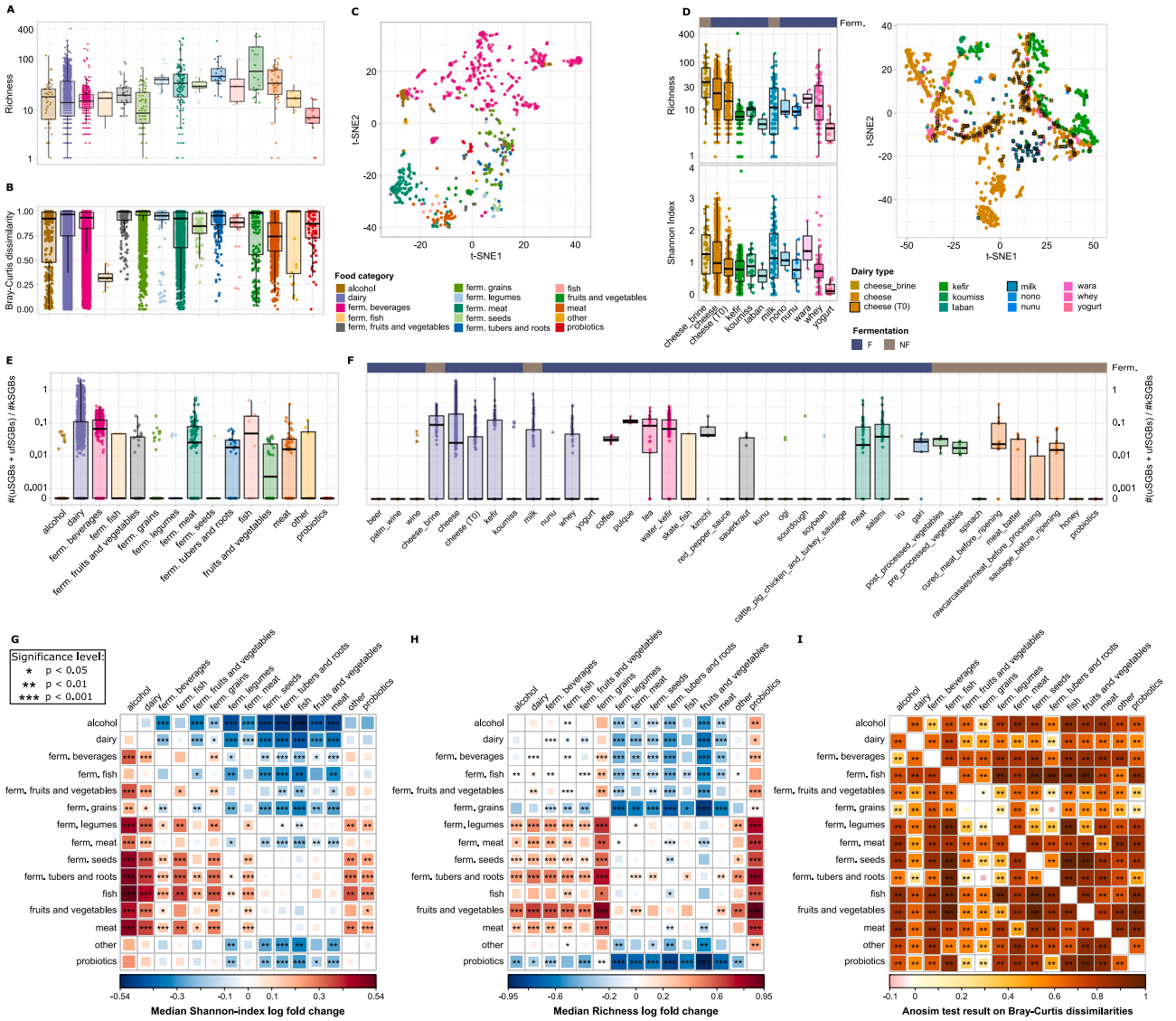


Figure S3. Taxonomic profiling enables characterization of food metagenomes across categories, related to Figure 4

(A) Differences in terms of alpha diversity (Richness calculated as the number of SGBs detected by MetaPhlAn) for the 2,533 food metagenomes grouped into 15 categories.

(B) Inter-category Bray-Curtis dissimilarity for the 15 categories.

(C) Beta diversity (t-SNE dimensionality reduction using Bray-Curtis distances) for food metagenomes, excluding dairy samples. Points are colored according to the food category.

(D–H) (D) Alpha (Richness and Shannon index) and beta (t-SNE dimensionality reduction using weighted UniFrac distances) diversity for the 1,650 dairy metagenomes grouped into different types (types with only one sample are not shown), with the indication of fermented (F) and non-fermented (NF) types. Samples belonging to cheese before ripening are highlighted as "cheese (T0)." The ratio between the number of uSGBs + ufSGBs and the number of kSGBs was computed for each sample and reported across (E) categories and (F) types (with $\geq$5 samples), with the indication of fermented (F) and non-fermented (NF) types. Samples belonging to cheese before ripening are highlighted as cheese (T0). Higher values are associated with higher unknownness levels. Statistical significance between categories in terms of alpha diversity for (G) Shannon index and (H) Richness. *p* values are computed through Wilcoxon-Mann-Whitney test and FDR correction.

(I) ANOSIM statistical test to assess differences between categories on Bray-Curtis distances. The heatmap is colored according to the R statistic.

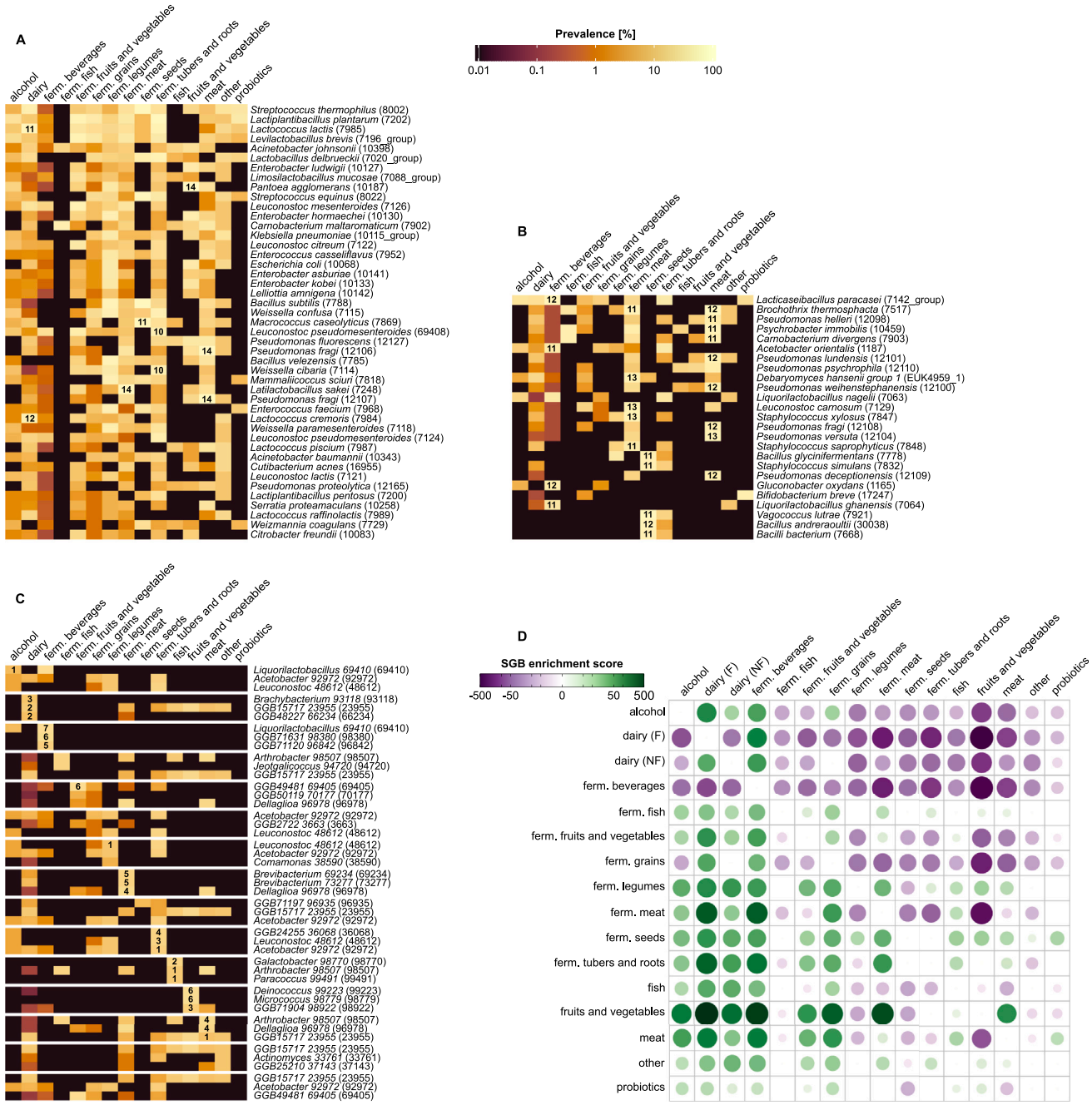


Figure S4. Taxonomic profiling enables sensitive detection of microbial species among food categories, related to Figure 4

(A) SGBs found in ≥9 categories. The heatmap is colored according to the prevalence in each category.

(B) SGBs most representative of food categories, i.e., with >10 significant comparisons, other than SGBs already shown in (A) and in Figure 4.

(C) Most representative uSGBs for each food category and their prevalence across food categories. In all these panels, written numbers represent the numbers of comparisons statistically significant between categories.

(D) The SGB enrichment score (see STAR Methods) for each pair of food categories of Figure 4F is reported here by dividing fermented (F) from non-fermented (NF) dairy. A score > 0 indicates a higher number of SGBs enriched in the row category.

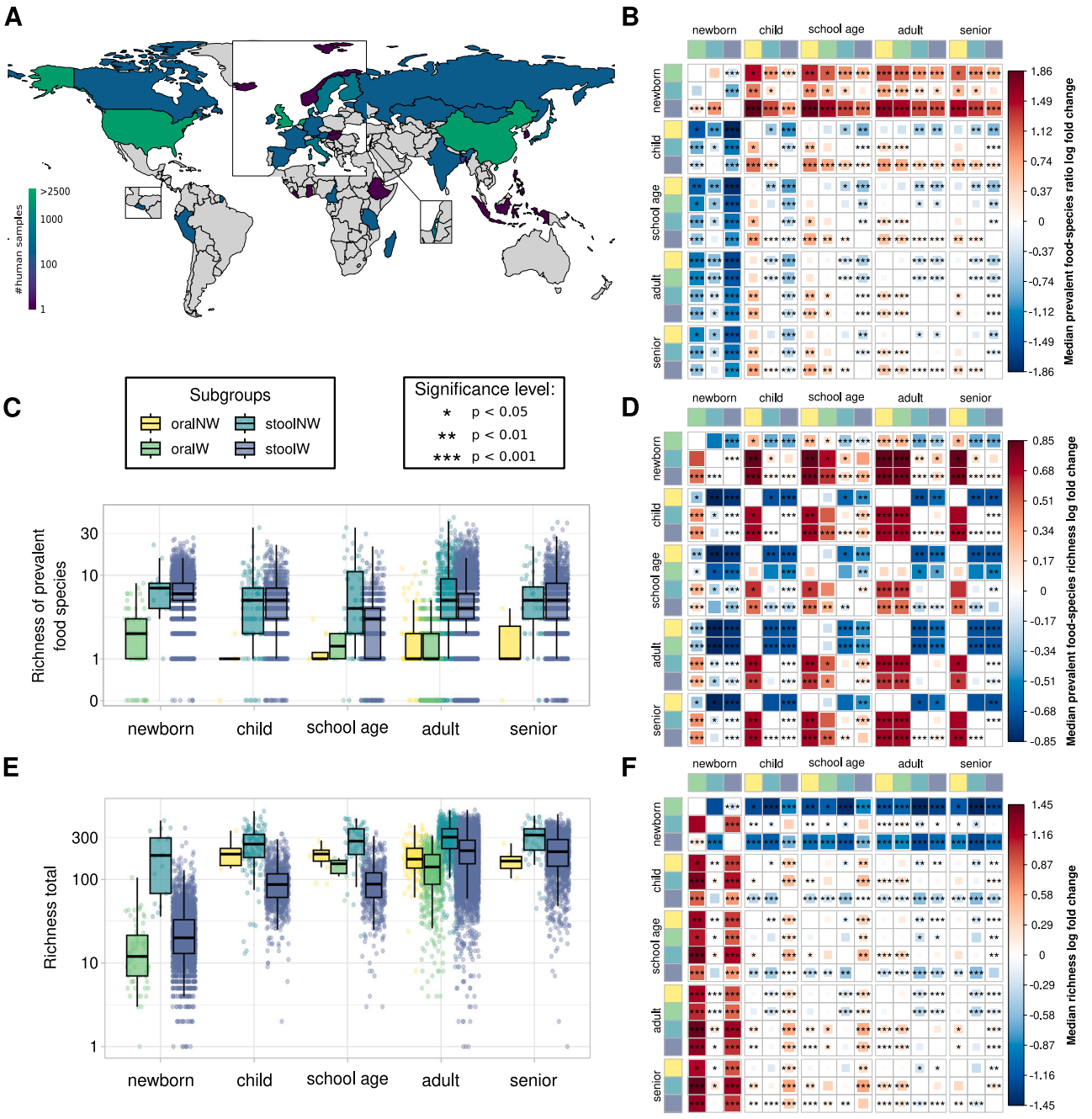
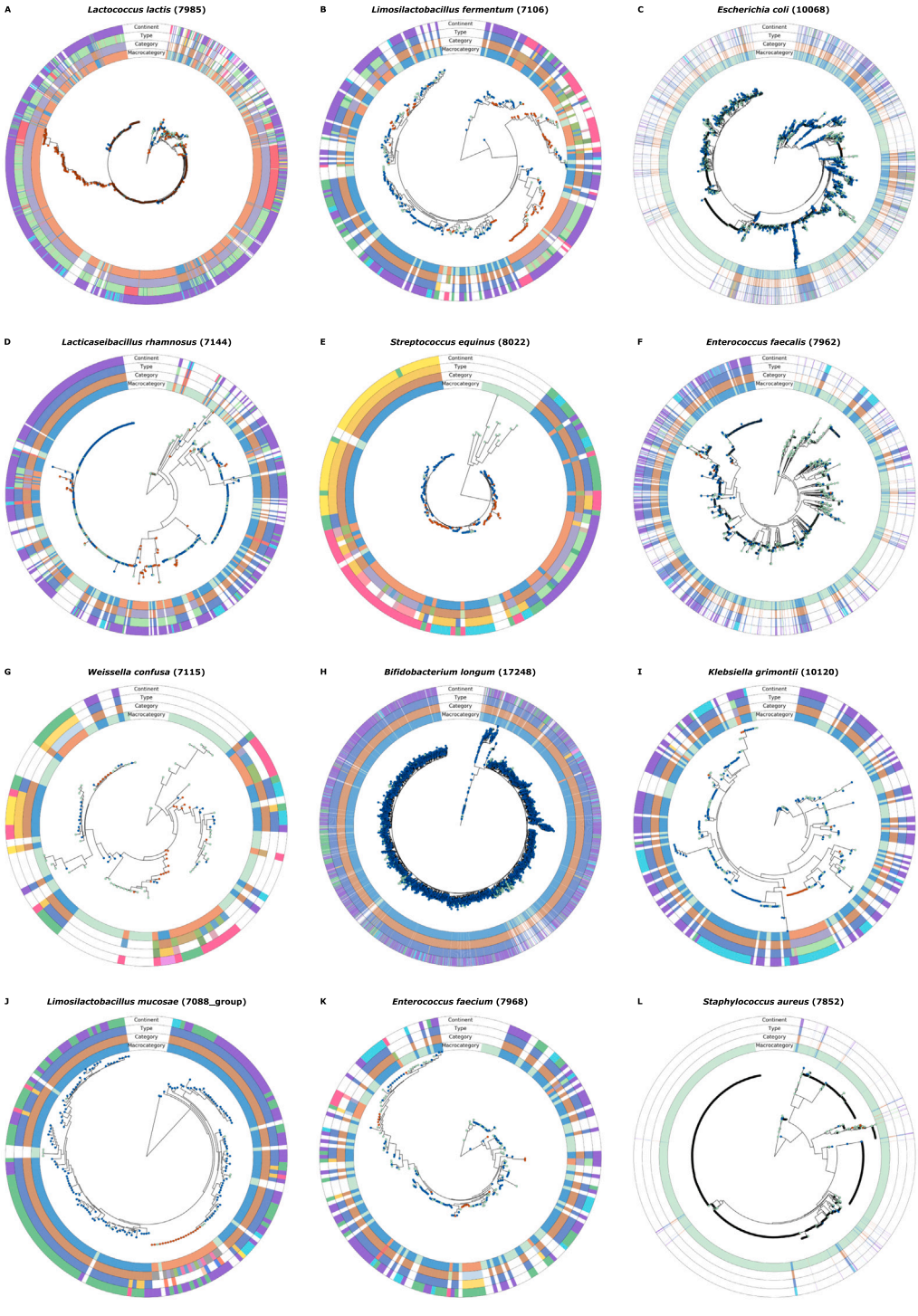


Figure S5. Overlaps in human microbiomes of SGBs prevalent in food, related to Figure 5

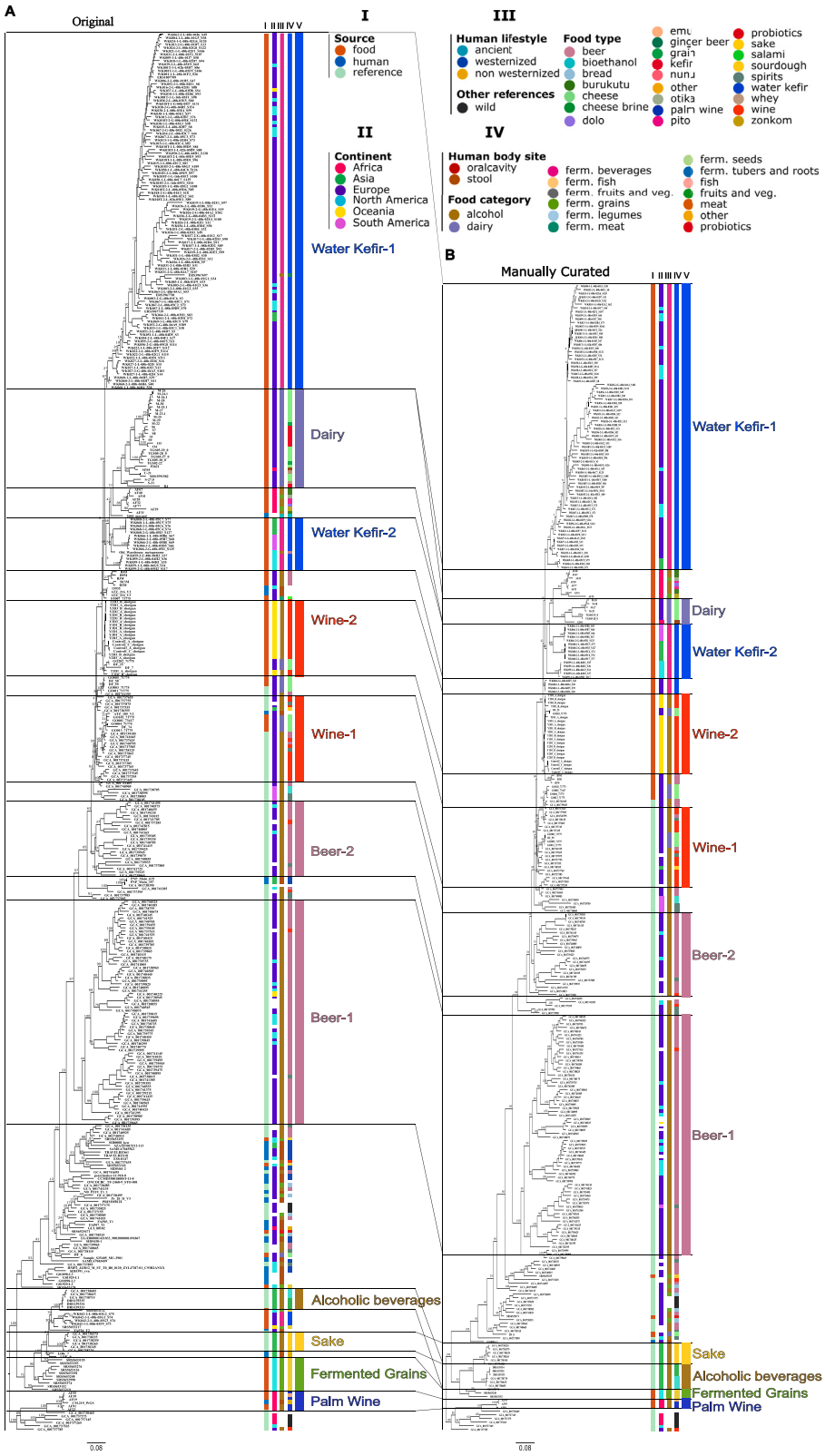
(A) Geographic distribution of the 19,833 publicly available human metagenomes considered in the analysis. Distribution of human samples in terms of (C) richness of prevalent food species and (E) total richness stratified by age category, body site, and lifestyle. Statistical significances (*p* values computed using Wilcoxon-Mann-Whitney test and FDR correction) of (C) and € are reported in (D) and (F), respectively; similarly (B) is relative to Figure 5B.



(legend on next page)

Figure S6. Additional phylogenetic trees of SGBs present in both food and human metagenomes, related to Figure 6

Trees were generated using the assembly free StrainPhlAn tool and by considering SGB-specific core genes for (A) *Lactococcus lactis* (SGB7985), (B) *Limosilactobacillus fermentum* (SGB7106), (C) *Escherichia coli* (SGB10068), (D) *Lacticaseibacillus rhamnosus* (SGB7144), (E) *Streptococcus equinus* (SGB8022), (F) *Enterococcus faecalis* (SGB7962), (G) *Weissella confusa* (SGB7115), (H) *Bifidobacterium longum* (SGB17248), (I) *Klebsiella grimontii* (SGB10120), (J) *Limosilactobacillus mucosae* (SGB7088_group), (K) *Enterococcus faecium* (SGB7968), and (L) *Streptococcus aureus* (SGB7852). Such trees complement the ones reported in Figure 6.



(legend on next page)

Figure S7. Phylogeny of eukaryotic *S. cerevisiae* by integrating metagenomes with isolate genomes, related to Figure 7

(A) Non-circular version of the phylogeny shown in Figure 7C in addition with node bootstrap values. The same phylogeny is shown in (B) after manual curation aiming at removing non-confidentially profiled strains (STAR Methods). In both trees, genome names are written on the leaves; metadata information is reported as color-coded columns (I–IV); clusters (V) are colored based on the majority of included food types and are named as proposed in literature or as defined by the newly acquired strains (for water Kefir-1 and water Kefir-2).