MICROBIAL GENOMICS

RESEARCH ARTICLE
Ksiezarek *et al., Microbial Genomics* 2022;8:000847
DOI 10.1099/mgen.0.000847

MICROBIOLOGY SOCIETY
OPEN DATA OPEN ACCESS

Genomic diversity of genus *Limosilactobacillus*

Magdalena Ksiezarek[1,2], Filipa Grosso[1,2], Teresa Gonçalves Ribeiro[1,2] and Luísa Peixe[1,2,*]

Abstract

The genus *Limosilactobacillus* (formerly *Lactobacillus*) contains multiple species considered to be adapted to vertebrates, yet their genomic diversity has not been explored. In this study, we performed comparative genomic analysis of *Limosilactobacillus* (22 species; 332 genomes) isolated from different niches, further focusing on human strains (11 species; 74 genomes) and their adaptation features to specific body sites. Phylogenomic analysis of *Limosilactobacillus* showed misidentification of some strains deposited in public databases and existence of putative novel *Limosilactobacillus* species. The pangenome analysis revealed a remarkable genomic diversity (only 1.3% of gene clusters are shared), and we did not observe a strong association of the accessory genome with different niches. The pangenome of *Limosilactobacillus reuteri* and *Limosilactobacillus fermentum* was open, suggesting that acquisition of genes is still occurring. Although most *Limosilactobacillus* were predicted as antibiotic susceptible (83%), acquired antibiotic-resistance genes were common in *L. reuteri* from food-producing animals. Genes related to lactic acid isoform production (>95%) and putative bacteriocins (70.2%) were identified in most *Limosilactobacillus* strains, while prophages (55.4%) and CRISPR-Cas systems (32.0%) were less prevalent. Among strains from human sources, several metabolic pathways were predicted as conserved and completed. Their accessory genome was highly variable and did not cluster according to different human body sites, with some exceptions (urogenital *Limosilactobacillus vaginalis*, *Limosilactobacillus portuensis*, *Limosilactobacillus urinaemulieris* and *Limosilactobacillus coleohominis* or gastrointestinal *Limosilactobacillus mucosae*). Moreover, we identified 12 Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologues that were significantly enriched in strains from particular body sites. We concluded that evolution of the highly diverse *Limosilactobacillus* is complex and not always related to niche or human body site origin.

DATA SUMMARY

The Supplementary Material associated with this article is available in the data repository Figshare under the link: https://doi.org/10.6084/m9.figshare.20052164 [1]

INTRODUCTION

For several years, the genus *Lactobacillus* (family *Lactobacillaceae*) has been under investigation aiming for proper classification of this bacterial group characterized by high levels of phenotypic and genotypic diversity [2]. In early 2020, the taxonomy of the genus *Lactobacillus* was revisited based on a polyphasic approach (core-genome phylogeny, pairwise average amino acid identity, clade-specific signature genes, physiological criteria and ecology), resulting in the reclassification of the genus *Lactobacillus* into 25 genera, including *Limosilactobacillus* gen. nov. (formerly the *Lactobacillus reuteri* group) [3].

To date, *Limosilactobacillus* comprises 23 validly published species (*Limosilactobacillus agrestis*, *Limosilactobacillus albertensis*, *Limosilactobacillus antri*, *Limosilactobacillus balticus*, *Limosilactobacillus caviae*, *Limosilactobacillus coleohominis*,

Received 06 October 2021; Accepted 16 May 2022; Published 15 July 2022
Author affiliations: [1]Laboratory of Microbiology, UCIBIO – Applied Molecular Biosciences Unit, REQUIMTE, Department of Biological Sciences, Faculty of Pharmacy, University of Porto, 4050-313 Porto, Portugal; [2]Associate Laboratory i4HB - Institute for Health and Bioeconomy, Faculty of Pharmacy, University of Porto, 4050-313 Porto, Portugal.
*Correspondence: Luísa Peixe, lpeixe@ff.up.pt
Keywords: animals; antimicrobial resistance; comparative genomics; human microbiome; *Lactobacillaceae*; pangenome.
Abbreviations: AMR, antimicrobial resistance; ANIb, average nucleotide identity based in BLAST+; CDS, coding sequence; COGs, Clusters of Orthologous Groups; CRISPR, clustered regularly interspaced short palindromic repeats; KEGG, Kyoto Encyclopedia of Genes and Genomes; KO, KEGG orthologue; NCBI, National Center for Biotechnology Information.
Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Eleven supplementary tables are available with the online version of this article.

Impact Statement

The recently proposed genus *Limosilactobacillus* (formerly included in *Lactobacillus*) comprises 23 species, some of which are well known for their importance in the food industry and probiotic potential. In this study, we applied comparative genomics approaches to investigate the differences between *Limosilactobacillus* species from different sources (human, animal and food), with particular focus on human isolates. The data presented here demonstrated that the genus *Limosilactobacillus* is highly diverse, comprising species that are present in multiple hosts and niches (e.g. *Limosilactobacillus fermentum*), those with more animal-specific adaptation (e.g. *Limosilactobacillus reuteri*) and putative novel *Limosilactobacillus* species not yet characterized. Although mostly antibiotic susceptible, *Limosilactobacillus* strains from food-producing animals are more prone to present acquired antibiotic-resistance genes. The common presence of bacteriocin-encoding regions, genes related to lactic acid production and CRISPR-Cas systems highlights that most *Limosilactobacillus* strains have defence mechanisms against other bacteria and/or foreign DNA. Additionally, to our knowledge, this is the first study to present comprehensive functional and metabolic predictions of human *Limosilactobacillus*, providing important insights for understanding the growth requirements and contribution of these bacteria to human nutrition and health.

Limosilactobacillus equigenerosi, *Limosilactobacillus fastidiosus*, *Limosilactobacillus fermentum*, *Limosilactobacillus frumenti*, *Limosilactobacillus gastricus*, *Limosilactobacillus gorillae*, *Limosilactobacillus ingluviei*, *Limosilactobacillus mucosae*, *Limosilactobacillus oris*, *Limosilactobacillus panis*, *Limosilactobacillus pontis*, *Limosilactobacillus portuensis*, *Limosilactobacillus reuteri*, *Limosilactobacillus rudii*, *Limosilactobacillus secaliphilus*, *Limosilactobacillus urinaemulieris*, *Limosilactobacillus vaginalis*), of which 7 were recently described [4, 5].

Limosilactobacillus members ferment a relatively broad spectrum of carbohydrates, yet several species do not ferment glucose [3]. Some *Limosilactobacillus* species, particularly *L. reuteri* and *L. fermentum*, are produced commercially for use as starter and probiotic cultures [6, 7]. All but four species (*L. fermentum*, *L. secaliphilus*, *L. urinaemulieris* and *L. portuensis*) were isolated from intestinal or faecal samples or were shown experimentally to have adapted to the intestine of vertebrate animals [3–5].

Advances in the healthy human microbiome reveal that some *Limosilactobacillus* species are often found in the gut, vagina and other human body niches, with *L. reuteri* being the most understood [7]. Of note, the healthy human microbiome appears to be colonized to a higher extent with *Lactobacillus* than *Limosilactobacillus* species [8, 9]. Nonetheless, it seems that in the urogenital tract, species belonging to *Limosilactobacillus* co-exist with *Lactobacillus* species [3, 9]. Thus, *Limosilactobacillus* species seem to be also relevant for human health and deserve as extensive a characterization as their *Lactobacillus* relatives.

The aim of this study was to evaluate the taxonomic diversity of the genus *Limosilactobacillus* and investigate the genomic diversity of *Limosilactobacillus* species from different sources (human, animal and food). We further performed comparative genomics of human isolates, including functional and metabolic characterization and niche-specific genomic content. To the best of our knowledge, this is the first study since the recent *Lactobacillaceae* reclassification focusing on the pangenome and characterizing the genomic diversity of the genus *Limosilactobacillus*.

METHODS

Genomes database

A total of 338 genomes representing complete and draft genomes of 22 *Limosilactobacillus* species, including the 2 recently published, i.e. *L. urinaemulieris* and *L. portuensis* [4], were retrieved from the National Center for Biotechnology Information (NCBI) Assembly database on 6th September 2020 (Table S1, available with the online version of this article). Six genomes were excluded from further analysis based on worse assembly statistics since there was better representative assembly available for the same strains, narrowing down the final collection to 332 *Limosilactobacillus* genomes (Table S1). Metadata related to downloaded assemblies was retrieved from the NCBI BioSample database. The genome sizes, guanine-cytosine content (G+C mol%), number of coding sequences (CDSs), tRNA and rRNA were extracted from Prokka v1.14.6 [10] and checkM v1.1.3 [11]. Draft genome completeness was accessed by checkM v1.1.3 [11] (Table S1).

Average nucleotide identity and phylogenomic analysis

Average nucleotide identity based in BLAST+ (ANIb) analysis on 332 assemblies was performed with pyani (v0.2) [12]. The ANIb results were interpreted according to widely established thresholds [13]. A percentage identity matrix was used to create a heatmap representing ANIb clusters by the R base heatmap package in R v4.0.3 [14] (Table S2). Phylogenomic analysis was performed using anvi'o v7.1 [15]. Single-copy core genes based on the Bacteria_71 collection from hidden Markov model profiles [16] were identified and 71 proteins were concatenated. FastTree version 2.1.11 [17] was used to recreate a maximum-likelihood phylogenomic tree with the Jones–Taylor–Thornton substitution model, local support of SH-like 1000 and "CAT" approximation (model that accounts for evolutionary rates across sites) with 20 rate categories. The resulting phylogenomic tree was edited in iTOL [18], while the ANIb figure was edited using Inkscape [19]. Genomes representing putative novel species were also submitted to the Type (Strain) Genome Server for genome-based taxonomy developed by the Leibniz Institute DSMZ (https://tygs.dsmz.de/).

Pangenome analysis

The anvi'o v7.1 [15] pipeline was used to profile hidden Markov models and predict genes using Prodigal [20]. Proteins were annotated with Clusters of Orthologous Groups (COGs; 2014 release) and Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologues (KOs) and KEGG modules [21] using anvi'o v7.1 (anvi'o tutorials, https://merenlab.org). Pangenome analysis was performed by anvi'o v7.1, using MUSCLE for sequence alignment [22], the Markov cluster algorithm [23] for clustering and NCBI BLASTP to assess amino acid sequence similarity. Partial gene calls were included in the analysis, and remaining parameters required to define gene clusters according to amino acid sequence homology were left as default [minbit heuristics of 0.5 and MCL (Markov Cluster algorithm) inflation of 2]. Gene collections were classified as follows: core genome included gene clusters present in 100% of genomes; accessory genome included three collections – softcore (gene clusters present in more than 95%), dispensable (gene clusters present in at least two genomes and in less than 95% of genome) and unique (singletons, genes present in just one unique genome).

The pangenome figure was visualized by anvi'o [15] and edited in Inkscape [19]. COGs categories distribution among strains was visualized in the R v4.0.3 [14] ggplot2 package v3.3.5 [24]. In the case of multiple COGs categories predicted per gene cluster, the most frequent was used as a representative, and in the case of multiple COGs categories predicted per gene, the first one was used as the most significant hit. Pangenome accumulation curves for species that had representation of at least 50 genomes was performed with Roary v3.13.0 [25].

Analysis of the accessory genome specific to isolation source was performed in R [14] v4.0.3 and the Venn diagram was created using VennDiagram R package v1.6.20 [26]. Only amino acid sequences with annotated COGs were included in this analysis. COGs specific to each origin group were identified with the VennDiagram package. The heatmap of accessory gene clusters for human strains was performed in R v4.0.3 [14] with phyloseq package v1.34.0 [27], hierarchical clustering was computed using Euclidean distance and the figure was edited in Inkscape [19].

Functional enrichment analysis based on enrichment scores [28] was performed in anvi'o. In short, functional enrichment analysis identifies functions (and/or KEGG modules) that might be associated only with a specific collection of genomes (anvi'o tutorials, https://merenlab.org). This analysis provides statistical support including Rao test, uncorrected *P* value and *q* value (*P* value adjusted for false detection rate) [28]. Only results with a *q* value less than 0.05 were considered significant and interpreted in this study. Additionally, metabolic reconstruction, including summary of KOs and estimation of pathway completeness using 75% as a threshold, was performed with anvi'o according to available tutorials (https://merenlab.org).

Other whole genome sequence analyses

Additional analyses included *in silico* bacteriocin-encoding gene prediction performed by BAGEL4 [29], detection of CDSs for Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR), CRISPR-associated genes (Cas) and spacers performed with CRISPRCasTyper v1.6.0 [30], and identification of prophage sequences performed by PhiSpy v4.2.19 [31]. Additionally, the detection of acquired genes mediating antimicrobial resistance (AMR) was performed by BLASTN from the NCBI BLAST2.8.1+ package [32] using a reference database extracted from ResFinder v4.1 [33].

RESULTS

Major genomic features

A total of 332 genomes representing 22 *Limosilactobacillus* species were analysed in this study (Tables 1 and S1). At the date of the genome retrieval, the NCBI Assembly database was lacking a representative genome for *L. caviae*. Most strains were isolated from animals (*n*=167 genomes, 50.3%), followed by humans (*n*=74, 22.3%) and food products (*n*=59, 17.8%). *L. reuteri* strains were the most sequenced, representing 60.2% of all available *Limosilactobacillus* genomes (Table 1).

Most of the available *Limosilactobacillus* genomes were draft genomes (83.4%; *n*=277/332 genomes; 15 species), and only 7 species (*L. fermentum*, *L. frumenti*, *L. gastricus*, *L. mucosae*, *L. pontis*, *L. reuteri*, *L. vaginalis*) had at least one complete genome (16.6%; *n*=55/332 genomes) (Table S1). Nevertheless, the mean completeness of draft genomes was 98.9%. The size of the genomes within *Limosilactobacillus* range from 1.49 Mbp (*L. reuteri* strain W1P28.032) to 2.65 Mbp (*L. reuteri* strain KLR1002), with a mean genome size of 2.08 Mbp. The mean number of CDSs was 2024, with *L. reuteri* W1P28.032 presenting the fewest CDSs (1388) and *L. fermentum* MTCC 8711 the most (2854). There is also a striking range in the G+C content, which ranges from 37.89 mol% (*L. agrestis* strain BG-MG3-A) to 53.45 mol% (*L. pontis* strain DSM 8475). The mean G+C content of genomes from human and food sources was similar (46.8 and 45.9 mol%, respectively), while for animals it was lower, 39.2 mol%. Moreover, the mean

Table 1. List of *Limosilactobacillus* species, number of genomes and origin available in NCBI database

Species	No. of genomes	Host
Limosilactobacillus agrestis	2	Rodents
Limosilactobacillus albertensis	2	Rodents, lemur
Limosilactobacillus antri	2	*Homo sapiens*
Limosilactobacillus balticus	2	Rodents, pheasant
Limosilactobacillus coleohominis	2	*Homo sapiens*
Limosilactobacillus equigenerosi	3	Horse
Limosilactobacillus fastidiosus	2	Rodents
Limosilactobacillus fermentum	79	*Homo sapiens*, cheese, fermented food products
Limosilactobacillus frumenti	3	Sourdough
Limosilactobacillus gastricus	3	*Homo sapiens*
Limosilactobacillus gorillae	1	Gorilla
Limosilactobacillus ingluviei	3	Pigeon
Limosilactobacillus mucosae	13	*Homo sapiens*, pig, boar, cattle
Limosilactobacillus oris	3	*Homo sapiens*
Limosilactobacillus panis	1	Sourdough
Limosilactobacillus pontis	3	*Homo sapiens*, sourdough
Limosilactobacillus portuensis	1	*Homo sapiens*
Limosilactobacillus reuteri	200	*Homo sapiens*, rodents, poultry, pig, cattle, horse, sheep, goat, badger, sourdough, dairy products
Limosilactobacillus rudii	2	Rodents
Limosilactobacillus secaliphilus	1	Unknown
Limosilactobacillus urinaemulieris	1	*Homo sapiens*
Limosilactobacillus vaginalis	3	*Homo sapiens*

genome size and number of CDSs were similar between *Limosilactobacillus* genomes from different origins (2.14 Mbp and 2086 CDSs in animals, 2.04 Mbp and 1980 CDSs in human, and 1.99 Mbp and 1952 CDSs in food).

Taxonomy

Analysis of ANIb showed clear *Limosilactobacillus* species separation based on the widely accepted threshold of 95% for species discrimination [13] (Fig. 1, Table S2). We observed near cut-off (94–95%) variations in ANIb values inside the *L. reuteri* clade supporting the existence of several subspecies as recently characterized by Li *et al.* [5]. Moreover, publicly available genomes of strains VA24_5, Lr4000, W1P44.042 and W1P28.032 deposited as *L. reuteri* and strain UMB0683 deposited as *L. pontis* should be reclassified since the ANIb values between these strains and the type strain of *L. reuteri* and *L. pontis* were all below 95%. The ANIb values between VA24_5, Lr4000 and *L. albertensis* Lr3000^T were >95% and between W1P44.042 and *L. mucosae* DSM 13345^T was 96% (Fig. 1, Table S2). Moreover, the ANIb value between strains W1P28.032 and UMB0683 was 81%, and between each strain and the closely related species *L. pontis* DSM 8475^T was 82.6 and 84%, respectively (Fig. 1, Table S2), suggesting that these strains may represent distinct and putative novel species.

The taxonomic position of strains was also elucidated by phylogenomic analysis based on 71 single-copy core proteins (Fig. 2). The results showed a clear distinction between genomes of different species. Moreover, strains VA24_5 and Lr4000 clustered with the type strain of *L. albertensis*, strain W1P44.042 grouped with the type strain of *L. mucosae*, and strains W1P28.032 and UMB0683 each formed an independent branch. Additionally, we submitted the two genomes from strains W1P28.032 and UMB0683 to the DSMZ Type (Strain) Genome Server, which predicted that both strains are putative novel species, with the closest relationship to *L. pontis* DSM 8475^T, with digital DNA-DNA hybridization (dDDH) values of 47.1 and 58.2%, respectively. On the basis of these analyses, we propose that strains VA24_5 and Lr4000 should be classified as members of the species *L. albertensis*, and strain W1P44.042 classified as a member of *L. mucosae*. Furthermore, two putative novel *Limosilactobacillus* species, here

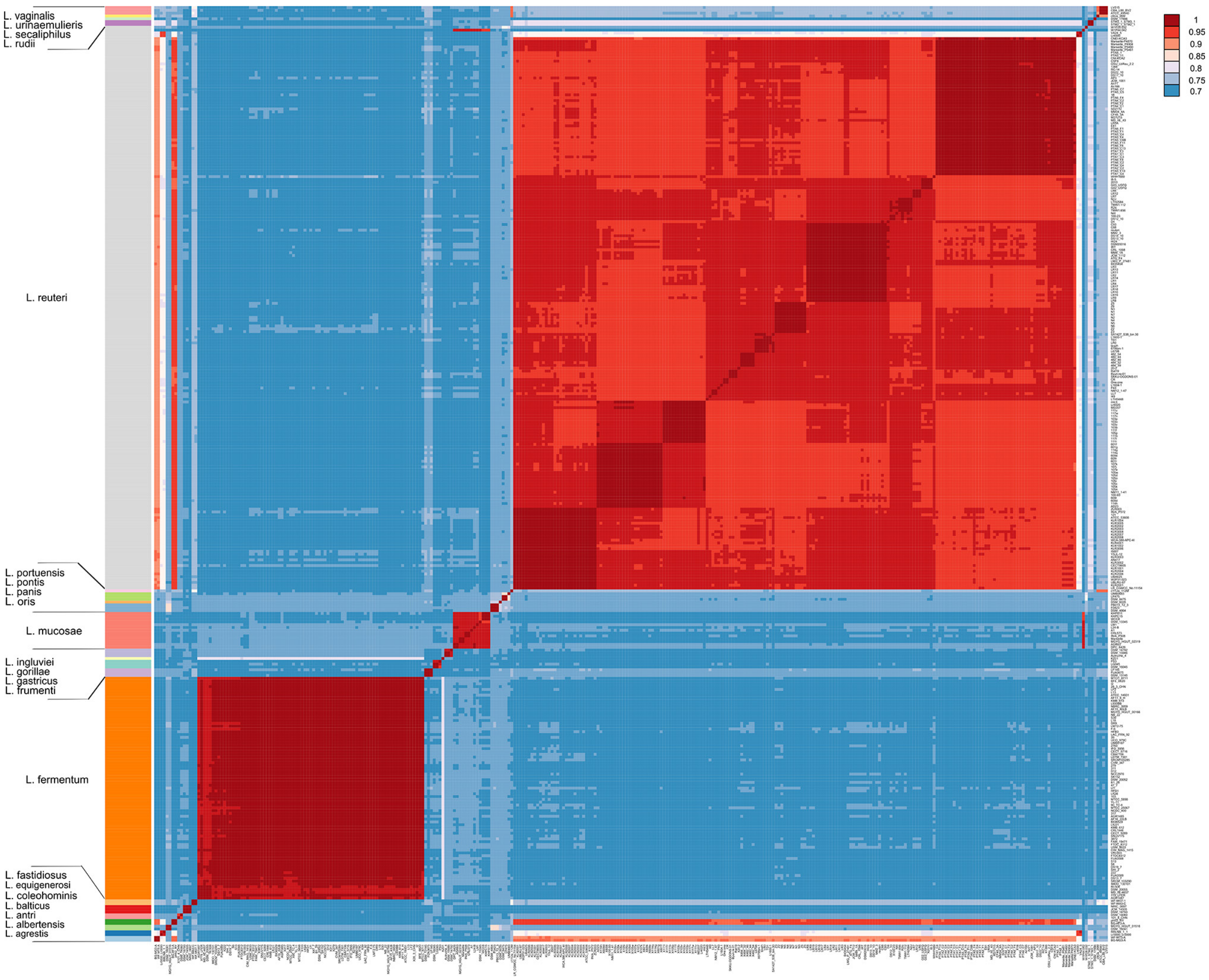

Fig. 1. Heatmap representing percentage identity of ANIb for 332 *Limosilactobacillus* genomes. The clusters of species according to the identification under which strains were deposited in the public NCBI database is provided on the left of the heatmap.

designated as *Limosilactobacillus* sp. nov. 1 (strain W1P28.032) and *Limosilactobacillus* sp. nov. 2 (UMB0683), were identified. *Limosilactobacillus* strains were not clustered by their host origin (Fig. 2), except for species with few representative genomes. The population structure of the 167 animal (mammals and birds) isolates was mostly associated with *L. reuteri*, and there were only four species (*L. reuteri*, *L. fermentum*, *L. mucosae* and *L. albertensis*) containing both human and animal isolates.

Pangenome of *Limosilactobacillus* spp.

The *Limosilactobacillus* pangenome (332 genomes, 22 species) was represented by 20 401 gene clusters (699 830 gene calls), of which only 39% (7937 gene clusters) had known COGs functions (Table S3). Only a minor part of the gene clusters presented metabolic predictions (21%, 4294 gene clusters, with known KOs, and 3%, 637 gene clusters, with known KEGG class). A high genomic variability was observed in this genome collection. The core genome contained only 266 gene clusters (1.3%; *n*=266/20401; 95 017 core genes), and accessory genomes included 20 135 gene clusters (98.7%; 2.8% softcore, 577 gene clusters, 202 378 genes; 57.2% dispensable, 11 659 gene clusters, 394 321 genes; 38.7% singletons, 7899 gene clusters, 8114 genes).

We analysed accumulation curves including core genes and the total number of genes for species that had more than 50 sequenced genomes i.e. *L. reuteri* (*n*=197 genomes) and *L. fermentum* (*n*=79) (Fig. S1). The pangenome accumulation curve did not reach a plateau for these two species, suggesting that the pangenome of *L. reuteri* and *L. fermentum* will continuously increase with new genomes. These findings indicate that these species have an open pangenome. In contrast, the core genome demonstrated a power trend line that plateaued. For the remaining species, only a limited number of genomes had been sequenced; thus, further sequencing is required to evaluate their genomic diversity.

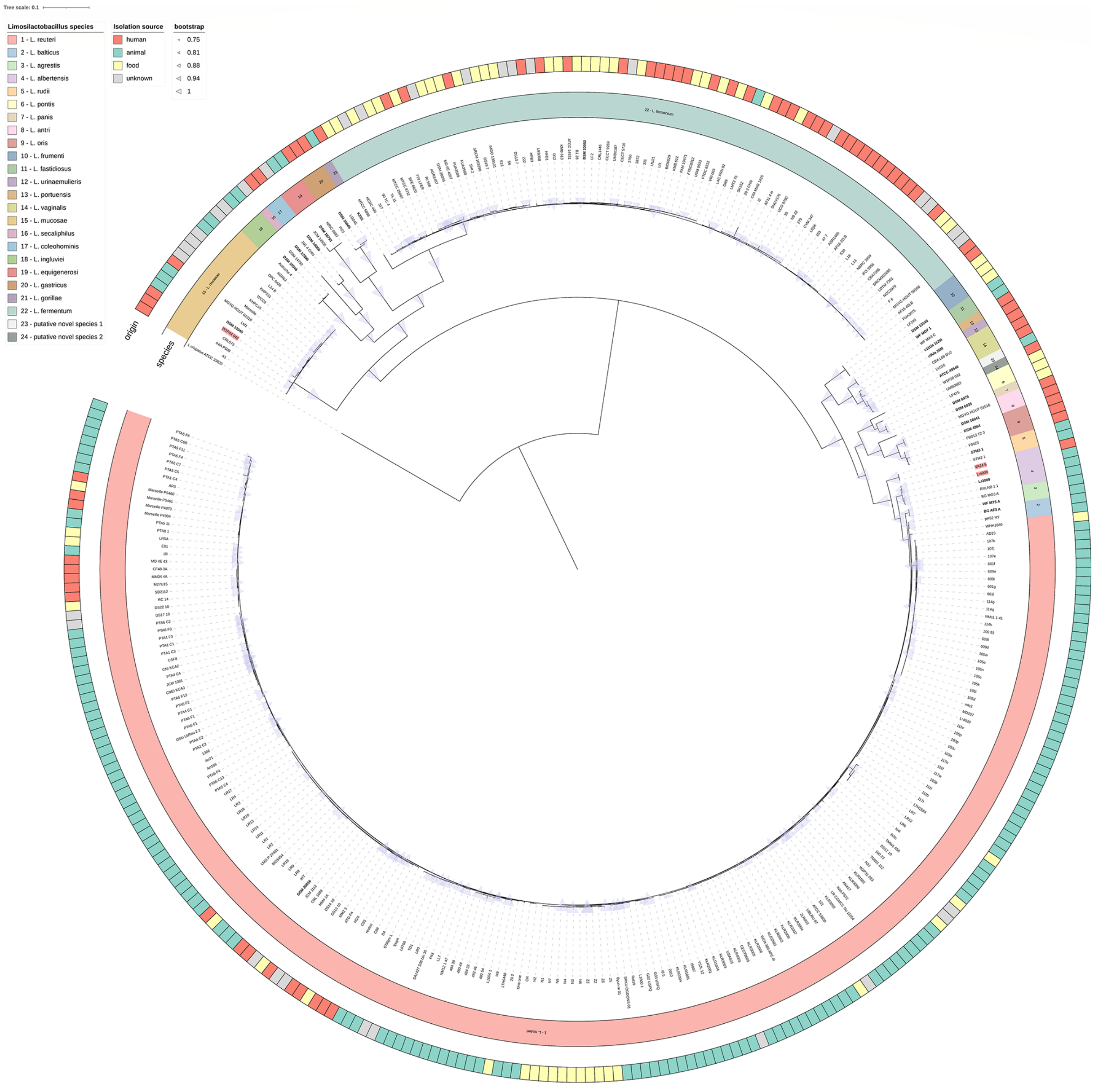

Fig. 2. Phylogenomic tree based on 71 single-copy core proteins created with the anvi'o 7.1 pipeline and edited in iTOL. The maximum-likelihood tree with Jones–Taylor–Thornton substitution model was built with FastTree v 2.1.11 with SH-like 1000 support. The tree includes 332 strains of which type strains identifiers are bold and misidentified species are marked with a red background. Additional data layers, i.e. species and strain origin, are incorporated in the figure according to the key. Bootstrap values are represented with a triangle symbol (only values ≥0.75 are shown). The scale bar represents 0.1 nucleotide substitutions per site.

Functional and metabolic characterization of *Limosilactobacillus* from different sources

As expected, core genes were mostly involved in housekeeping processes, including metabolic functions (e.g. COGs categories E, F, G, H, C), and non-metabolic-related central processes such as ribosomal biogenesis (category J) and replication (L) (Table S3). Similar functions were commonly observed within softcore genes. Regarding the accessory genome, the largest number of genes was related to translation, ribosomal structure and biogenesis (category J), while singletons were commonly associated with carbohydrate metabolism (category G), cell wall biogenesis (category M) and transcription (category K). Of note, most strains encode D-lactate dehydrogenase (*n*=329) and L-lactate dehydrogenase (*n*=331), which are associated with lactic acid production (Table S3).

To investigate whether surviving in different niches requires special features, we analysed the accessory genome of *Limosilactobacillus* with respect to the isolation source, including only gene calls that had annotated COGs (*n*=261 256 gene calls; 2116 unique

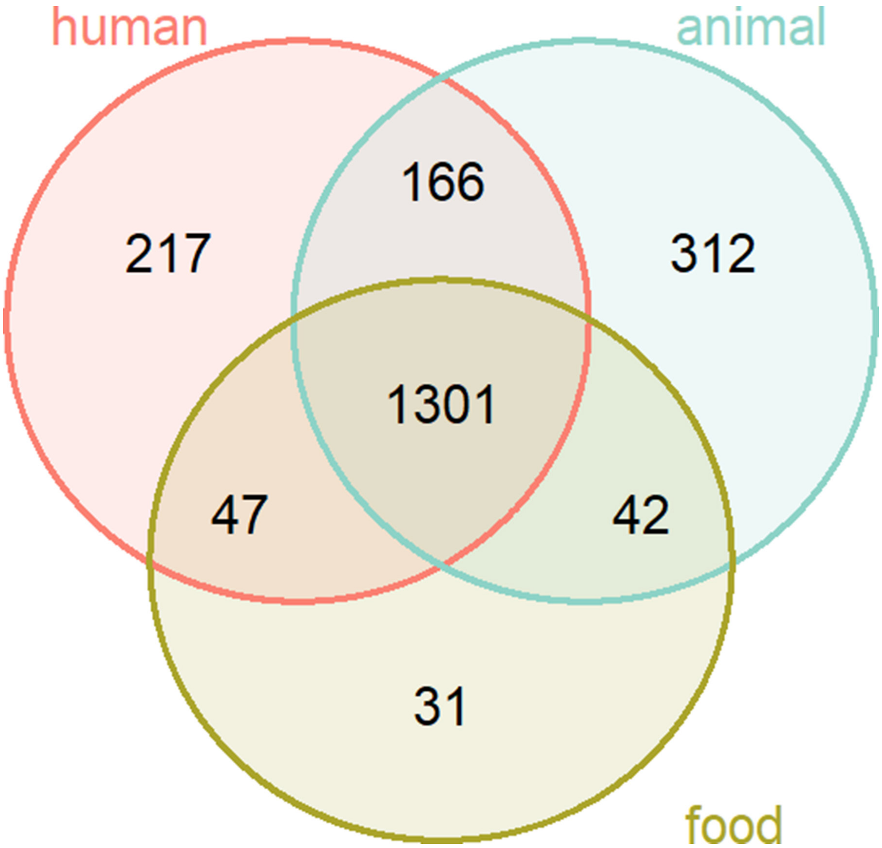

Fig. 3. VennDiagram image representing the number of accessory genes with annotated COGs functions (*n*=261 256 gene calls; 2116 unique COGs) that are shared and/or unique between *Limosilactobacillus* strains of different origin (i.e. human, animal, food).

COGs). We found that 1301 COGs are shared between strains independently of their origin (Fig. 3, Tables S1 and S3), with many of them corresponding to the softcore group (508 COGs functions), such as ribosomal proteins (category J), translation elongation factors (category J), aminotransferases (category E) or various genes related to carbohydrate metabolism (category G). We also found 560 niche-specific COGs (312 from animals, 217 from humans and 31 from food), and 255 dual-host shared COGs. The most prevalent COGs function unique to human isolates was phosphoribosyl-ATP pyrophosphohydrolase involved in amino acid metabolism (54%; 40/74 human strains), while *N*-acetylmuramoyl-L-alanine amidase CwlA, involved in cell wall organization, was the most frequent in animal isolates (21%; 35/167 animal strains), and the restriction-modification system DNA methylase subunit related to defence mechanisms was the most prevalent in food isolates (12%; 7/59 food strains). However, none of the COGs was present in all strains isolated from a particular niche.

Metabolic enrichment was performed according to strain isolation source, i.e. identification of KOs that were more frequently associated with a particular origin (i.e. isolated from animals, humans or food products) (Table S4). Overall, strains isolated from humans were enriched in KOs associated with galactose (K02773, K02774), glycogen (K00688), starch and sucrose (K00703), and sulfur (K15554, K15553) metabolism, folate catabolism (K12941), the secretion system (K12063), and tetracycline antibiotic resistance (K08151). Strains isolated from animals were enriched in KOs associated with methane (K05884, K01007) and carbohydrate (e.g. K01624, K01625, K01686, K01685), and amino acid (K00558) metabolism, lipoic acid (K16869), cobalamin (e.g. K02224, K03394, K05934, K05895, K05936, K06042), and haem (e.g., K01772, K01698, K01749, K01845) biosynthesis, and tetracycline resistance (K08168). Strains isolated from food were enriched in KOs associated with amino acid metabolism (K00383), lipid biosynthesis proteins (K00142), KDP operon response regulator KdpE (K07667), ArsR family transcriptional regulator (K21903) and membrane transport (K10117). Detailed information on KOs and KEGG modules associated with strains of particular origin is provided in Table S4.

Limosilactobacillus from humans

The 74 *Limosilactobacillus* genomes from human strains belonged to 11 species, including *L. albertensis*, *L. antri*, *L. coleohominis*, *L. fermentum*, *L. gastricus*, *L. mucosae*, *L. oris*, *L. portuensis*, *L. reuteri*, *L. urinaemulieris* and *L. vaginalis*, and *Limosilactobacillus* sp.

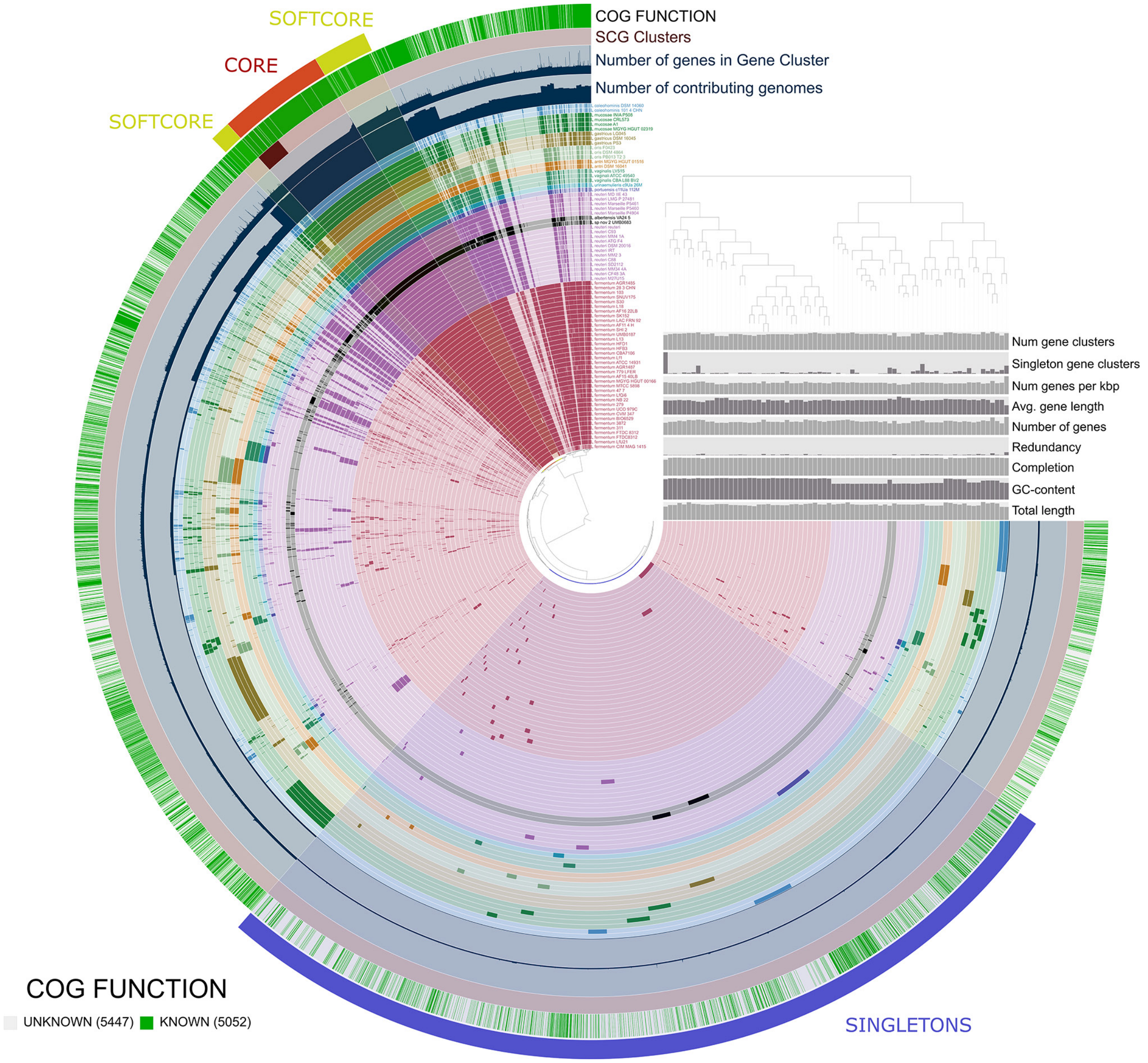

Fig. 4. Pangenome of all *Limosilactobacillus* strains isolated from human hosts (11 species, 1 putative novel species, 74 genomes) generated by anvi'o. Genomes are organized based on the tree of frequencies of gene clusters (top right). Each colour represents different species. External rings incorporate additional data regarding single-copy genes (SCG) clusters, number of genes per gene cluster and number of contributing genomes. The external white–green ring represents COGs functionality annotation, with green standing for known and white for unknown functions. Outside highlights represent particular gene collections: core, red; softcore, yellow; and singletons, blue. Additional information such as total length, G+C content, completion, redundancy, number of genes, mean gene length, number of genes per kbp, singleton gene clusters and number of gene clusters are represented by bars at the top right.

nov. 2 (strain UMB0683) (Table S5). The highest number of genomes belonged to *L. fermentum* (*n*=36) and *L. reuteri* (*n*=17), while remaining species had relatively poor representation (1–4 genomes each). Most strains were isolated from the gastrointestinal tract (*n*=31), followed by urogenital sources (*n*=9 vagina, *n*=7 urine), the oral cavity (*n*=6) or breast milk (*n*=5). For 16 strains, the human body site was not reported (Table S1).

The pangenome of human *Limosilactobacillus* spp. was represented by 10 499 gene clusters (151 749 genes), with 5052 clusters of known COGs function (Fig. 4). Human isolates demonstrated high genomic variability, since the core genome was represented only by 453 gene clusters, which correspond to 4.3% of all gene clusters (37 033 genes, 24.4%) (Fig. 4). Within the accessory genome (10 046 gene clusters, 95.7%; 114 716 genes, 75.6%), the softcore group was represented by 316 gene clusters (24 553 genes), the dispensable genome had 5967 gene clusters (86 215 genes) and singletons comprised 3763 clusters (3948 genes). Most of the gene clusters in the core genome have known COGs function (97.8%), contrary to the accessory genome where most have unknown COGs function (54%).

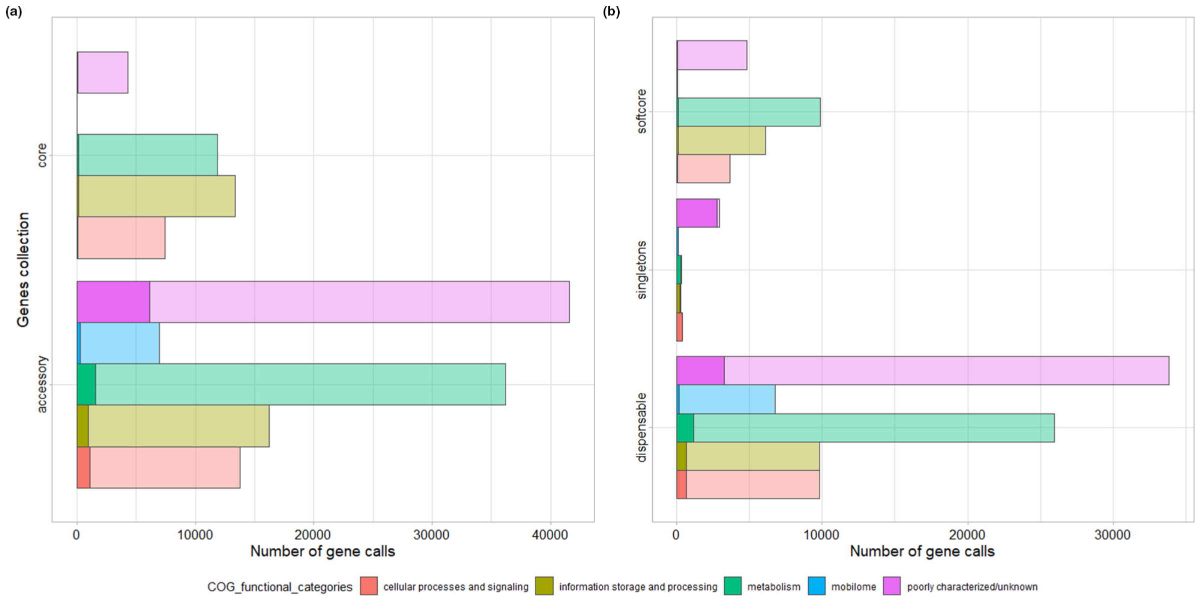

Fig. 5. Pangenome-specific gene collections grouped by COGs functional categories. Specific categories included: category J, K, L into information storage and processing; categories D, V, T, M, N, U, O into cellular processes and signalling; categories C, G, E, F, H, I, P, Q into metabolism; category X into mobilome; categories R, S and not classified are designated as poorly characterized/unknown. The transparent bars represent the number of gene calls, while non-transparent bars represent the number of gene clusters. In the case of multiple COGs categories predicted per gene, the first one was used as the most significant hit. In the case of multiple COGs categories predicted per gene cluster, the most frequent was used as a representative. (a) Core and accessory genome grouped by COGs functional categories. (b) Accessory genome represented by softcore, dispensable and singleton genes.

Most of the core genes were involved in information, storage and processing [translation, ribosomal structure and biogenesis (category J, 8236 genes), replication, recombination and repair (category L, 3183 genes)], and metabolism [nucleotide transport and metabolism (category F, 3172 genes), amino acid transport and metabolism (category E, 2440 genes) and carbohydrate transport and metabolism (category G, 2311 genes)]. Nearly the same functional pattern was observed for softcore genes compared with core genes (Fig. 5a, b). Many dispensable genes were related to metabolism (amino acids, 7492 genes; carbohydrates, 4520 genes; and coenzyme transport and metabolism, 4032 genes), followed by those related to cell wall biogenesis (category M, 3243 genes) and defence mechanisms (category V, 2293 genes), transcription (category K, 4983 genes), replication (category L, 3556 genes) and translation (category J, 1723 genes), and those related to mobilome (category X, 7942 genes). Singletons with a known function were associated with cell wall/membrane/envelope biogenesis (category M, 130 genes), metabolism of carbohydrates (category G, 113 genes), transcription (category K, 110 genes), replication, recombination and repair (category L, 93 genes) and mobilome – prophages, transposons (category X, 86 genes).

We identified 15 and 4 gene clusters including duplicated genes (estimation of 2 genes in the gene cluster/strain) in the core and softcore genome, respectively. The majority of duplicated genes in the core were involved in amino acids (COGs category E, e.g. ABC-type polar amino acid transport system, ATPase component), followed by carbohydrates (COGs category G, e.g. Na^+/melibiose symporter or related transporter) metabolism, and genes related to cell wall biogenesis (COGs category M, e.g. LysM repeat). For the softcore genome, the duplicated genes were mostly involved in energy production and conversion (category C, e.g. malate/lactate dehydrogenase). A detailed list of genes within each gene cluster, their functionality and homogeneity indices is provided in Table S6.

Metabolic reconstruction revealed that all human *Limosilactobacillus* spp. strains were predicted to utilize galactose (Leloir pathway), synthesize UDP-*N*-acetyl-D-glucosamine and had F-type ATPase for energy metabolism. Also, the majority of strains use the pentose phosphate pathway and/or the Embden–Meyerhof pathway for carbohydrate metabolism, biosynthesis of arginine, ornithine, proline, cysteine, methionine, lysine and threonine for amino acid metabolism, the phosphate acetyltransferase-acetate kinase pathway associated with carbon fixation, and complete pathways for coenzyme A, thiamine, tetrahydrofolate, riboflavin and molybdenum cofactor biosynthesis (Table S7).

To understand whether human strains cluster according to their isolation source in the human body, we performed clustering based on Euclidean distance using their accessory genome (Fig. 6). This comparison was based on the presence/absence of softcore, dispensable and singleton gene clusters. Overall, the distribution of these 10 046 gene clusters was different in the human strains, showing that the accessory gene clusters were conserved within species. This clustering showed that 74 human *Limosilactobacillus*

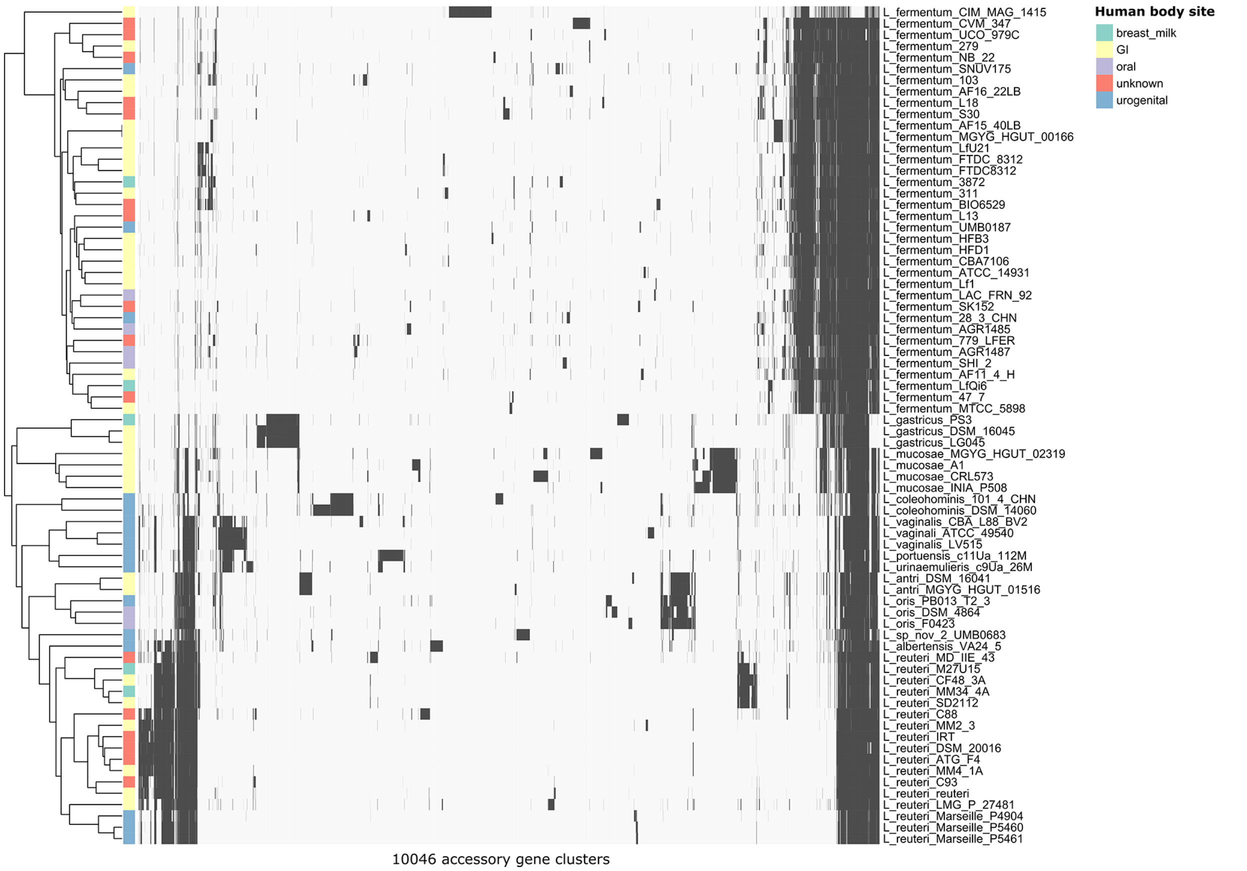

Fig. 6. Heatmap representing 10 046 accessory gene clusters (presence/absence) of 74 human *Limosilactobacillus* strains. The dendrogram on the left represents hierarchical clustering based on Euclidean distance. Additional information regarding the isolation site within a human body is presented in the bar on the left of the heatmap. GI, Gastrointestinal.

strains cluster independently of their human isolation source, with some exceptions (e.g. urogenital *L. coleohominis*, *L. vaginalis*, *L. portuensis*, *L. urinaemulieris* and gastrointestinal *L. mucosae*), for which only a few genomes were available. Of note, some human strains belonging to *L. reuteri* clustered according to their origin (e.g. urogenital strains); however, further observations are not possible due to the presence of strains of unknown origin (e.g. cluster of strains from gastrointestinal origin intermingled with those of unknown origin).

To further clarify biological functions of human-associated *Limosilactobacillus*, we first performed COGs enrichment analysis in strains for which the isolation source was detailed, although few genomes were available for some body sites (e.g. six oral strains). Five COGs were found to be significantly enriched (q value <0.05) in oral or oral and urogenital strains. The enriched COGs for oral strains were associated with transcription (COG4977, COGs category K) and carbohydrate metabolism (COG3533, COGs category G), while those identified in oral and urogenital strains were more enriched in information, storage and signalling (COG2231, COGs category L; COG3695, category K) and cellular processes and signalling (COG3290, COGs category T).

Subsequently, we compared the enriched KOs to explore the difference between the *Limosilactobacillus* strains in terms of their biological capabilities and to highlight adaptation to particular human sources (only entries with q value below 0.05 were considered) (Table 2). Some oral-associated *Limosilactobacillus* genomes uniquely presented categories related to metabolism (i.e. glycosidases) that were absent in other human sources, whereas some urogenital-associated genomes only presented categories related to defence mechanisms (i.e. toxin ParE1/3/4), signal transduction (i.e. sensor histidine kinase YcbA), and fructoselysine degradation pathway (i.e. fructoselysine 6-phosphate deglycase and fructoselysine 6-kinase).

AMR determinants, bacteriocins, CRISPR-Cas system and phages in *Limosilactobacillus* spp.

The majority of *Limosilactobacillus* spp. were predicted as antibiotic susceptible (83%, n=268/322). Nonetheless, 18 AMR genes were identified among the remaining 54 strains (Table S8), with tetracycline [*tet*(L), *tet*(W), *tet*(M), *tet*(C), *tet*(O/W)], aminoglycoside [*aadD*, *aph(3')-III*, *ant(6)-Ia*, *ant(9)-Ia*], macrolide-lincosamide-streptogramin [*erm(A)*, *erm(B)*] and

Table 2. A list of 12 KOs associated with particular isolation sources (only strains with a known human isolation source were considered)

Only KOs found as significant were included (*q* value <0.05). UGT, Urogenital tract; GI, Gastrointestinal.

Source	KO accession no.	Enzyme entry	KO family	Detection in strains from specific human body site (%)			
				UGT	Oral cavity	GI	Breast milk
Oral	K18205	EC:3.2.1.185	Non-reducing end β-l-arabinofuranosidase	0	33.33	0	0
	K12983	EC:2.4.1.-	UDP-glucose:(glucosyl)LPS β-1,3-glucosyltransferase	18.75	100	19.35	20
UGT, oral	K07457	–	Endonuclease III related protein	62.5	33.33	6.45	0
	K07443	–	Methylated-DNA-protein-cysteine methyltransferase related protein	62.5	33.33	6.45	0
UGT	K19092	–	Toxin ParE1/3/4	56.25	0	0	0
	K19802	EC:5.1.1.20	l-Ala-d/l-Glu epimerase	37.5	0	0	0
	K07717	EC:2.7.13.3	Two-component system, sensor histidine kinase YcbA	37.5	0	0	0
	K01751	EC:4.3.1.15	Diaminopropionate ammonia-lyase	31.25	0	0	0
	K10708	EC:3.5.-.-	Fructoselysine 6-phosphate deglycase	31.25	0	0	0
	K11382	–	MFS transporter, OPA family, phosphoglycerate transporter protein	31.25	0	0	0
	K00407	–	Cytochrome c oxidase cbb3-type subunit IV	31.25	0	0	0
	K10710	EC:2.7.1.218	Fructoselysine 6-kinase	31.25	0	0	0

lincomycin [*lnu*(A)] resistance genes being commonly identified. Moreover, AMR genes were mostly identified in *L. reuteri* isolated from animals (72%, *n*=39/54, mainly from chickens and pigs). In human strains, only genes conferring resistance to tetracyclines [*tet*(W), *tet*(C)] and/or to lincomycin [*lnu*(A)] were identified. Interestingly, AMR genes from at least three antimicrobial classes, consistent with a multidrug-resistant genotype, were identified in 11 strains (*L. reuteri* strains, isolated from animals).

The *in silico* analysis of putative bacteriocin gene clusters (Table S9) showed that 233 strains (70.2%; 17 *Limosilactobacillus* species) were potential bacteriocin producers. In total, 396 putative bacteriocins were predicted and identified as: enterolysin A (class III; *n*=318; 15 species), sactipeptides (class I; *n*=48; *L. reuteri*, *L. fermentum*, *L. panis*, *L. antri*), gassericin K7B (class II; *n*=23; *L. reuteri*, *L. albertensis*), carnolysin (class I; *n*=4; *L. reuteri*, *L. gastricus*), gassericin T (class II; *n*=1 *L. albertensis*), acidocin 8912 (class II; *n*=1 *L. fermentum*) and cytolysin ClyLs (class I; *n*=1 *L. gastricus*). Forty-nine strains, mostly *L. reuteri* from animals, encoded more than one bacteriocin. Overall, the bacteriocins were mostly identified in *Limosilactobacillus* strains from animal sources (*n*=143/332), covered all three classes, yet class III was the most detected.

Six types of CRISPR-Cas systems (I-C, I-E, I-G, II-A, III-A and III-D) were identified in 106 *Limosilactobacillus* strains (32%), of which 41 had more than one type (Table S10). The type II-A was the most frequent (*n*=73 strains), followed by I-E (*n*=46 strains) and III-A (*n*=14 strains). Interestingly, in strains belonging to *L. fermentum*, five different CRISPR-Cas types were observed, while *L. reuteri* had only type II-A. Overall, CRISPR-Cas-positive strains were common in *L. fermentum* (60 isolates; 76%), while in *L. reuteri* they were much less frequent (26 isolates; 13%). Regarding isolation source, CRISPR-Cas systems were more frequent among *Limosilactobacillus* isolated from humans (32%; *n*=34/106), followed by food (26.4%; *n*=28/106) and animals (25.5%; *n*=27/106). CRISPR-Cas types I-C, I-E, II-A and III-A were detected in strains from humans, animals and food origins, while I-G and III-D were observed each in a human isolate. Among the *Limosilactobacillus* strains harbouring CRISPR-Cas systems, 2 to 111 spacers (median 20) were identified that can target foreign DNA. However, the nucleic acid source of the majority of spacer sequences remains unknown as no positive hits were identified.

We identified complete or partial putative prophages in 55.4% (*n*=184/332) *Limosilactobacillus* genomes (Table S11). The lengths of these prophage elements ranged from 4.9 to 72.1 kb (mean 20.5 kb). Fifty-five strains had two to four prophages. Moreover, the prevalence of predicted prophages varied considerably among strains collected from different sources. The highest prevalence was observed in animal strains (*n*=110; 1–4 prophages), followed by food (*n*=29; 1–3 prophages) and humans (*n*=30; 1–2 prophages). Most predicted phage genes were uncharacterized proteins (89.4%; *n*=5425/6066 genes), and various mobilome-associated genes, e.g. tyrosine recombinases, transposases (e.g. IS*3*, IS*30*, IS*200*/IS*605*).

DISCUSSION

This study is, to the best of our knowledge, the first to present a comprehensive pangenome analysis of *Limosilactobacillus*, including the largest number of *Limosilactobacillus* genomes (*n*=332) and species (*n*=22) retrieved from publicly available repositories. The number of complete genomes is still low and reserved mostly to species with well-explored beneficial activities, such as *L. reuteri* and *L. fermentum* [34], or/and isolated from animals. Thus, comprehensive detection of certain genomic characteristics associated with particular species or host-/niche-specificity can be challenging.

The mean genome size of *Limosilactobacillus* was 2.08 Mbp with an average G+C content of 42.75 mol%, which was consistent with data from Zheng and colleagues [3]. Overall, ANIb and phylogenomic analysis revealed that public databases still contain misidentified strains (VA24_5 and Lr4000 here identified as *L. albertensis* and W1P44.042 as *L. mucosae*), and strains (W1P28.032 and UMB0683) that should be further characterized to confirm their potential as novel members of genus *Limosilactobacillus* (Fig. 2).

Previous comparative genomics studies included members of *Limosilactobacillus* (*L. reuteri*, *L. fermentum*, *L. mucosae* and *L. oris*) and other related genera (former *Lactobacillus*; *Lactiplantibacillus*, *Latilactobacillus*, *Lacticaseibacillus*, *Levilactobacillus*, *Lentilactobacillus*, *Companilactobacillus*, *Paucilactobacillus*, *Apilactobacillus*, *Ligilactobacillus*, *Fructilactobacillus*) [35, 36]; however, a pangenome analysis comprising all *Limosilactobacillus* species had not been performed yet. Our findings revealed that the core genome of *Limosilactobacillus* spp. is relatively small (1.3%), demonstrating the high genomic diversity of this genus. Additionally, our findings support that *L. reuteri* and *L. fermentum* have an open pangenome, as previously noted [37, 38]. An open pangenome is characteristic of species for which genomic content is still not completely defined, and its diversity increases with constant acquisition of genes [39].

It is noteworthy that our analyses showed that strains from humans, animal and food cannot be differentiated based on core phylogenomic analysis (Fig. 2); however, within the accessory genome, a subset of functions were uniquely present in some strains from each origin, which might enable their better adaptation to these niches. In fact, pangenome characterization showed that members of *Limosilactobacillus* shared little genomic similarity (98.7% of pangenome comprises accessory gene clusters), suggestive of a diverse gene pool, customized more to suit species-specific instead of niche-specific needs. Nonetheless, a more conclusive view regarding the *Limosilactobacillus* genomic heterogeneity can be obtained only after including more genomes for members with few representations.

While the majority of *Limosilactobacillus* genomic studies have focused on species most relevant to industry, e.g. the probiotic market [40–43], we focused our analysis on human *Limosilactobacillus*, since several species were identified in different human body sites, and occasionally in high relative abundance. For instance, *L. vaginalis* and *L. urinaemulieris* were found in high relative abundance in the vaginal microbiome of healthy women [44; M. Ksiezarek and others, unpublished data], *L. mucosae* in the vagina of Indian women and the faeces of healthy children [45], *L. fermentum* in breast milk and in saliva from patients with dental caries [46, 47], and *L. reuteri* in the intestinal microbiota of obese adults [48].

We reported here that 4.3% of human *Limosilactobacillus* pan-gene clusters correspond to the core genome, which is higher than previously reported for other *Lactobacillaceae* inter-species studies (up to 2.8%) [35, 49, 50]. Analysis of the core genome supports that fundamental processes such as translation, ribosomal structure and biogenesis, replication, recombination and repair, and information storage and processing were conserved and essential for bacterial survival, similarly to previous observations [36, 51, 52]. Of note, we observed duplications for core genes related to, for example, amino acid and carbohydrates metabolism, which seems to play a significant role in the biological evolution [53].

We also identified predicted *in silico* conserved metabolic pathways for human *Limosilactobacillus* spp. including sugar metabolism (pentose phosphate pathway and Embden–Meyerhof pathway), energy metabolism (F-type ATPase), amino acids metabolism (arginine, ornithine, proline, cysteine, methionine, lysine and threonine biosynthesis) or metabolism of cofactors and vitamins (coenzyme A, thiamine, tetrahydrofolate, riboflavin and molybdenum cofactor), which could be of interest to understand growth requirements and contribution of these bacteria to human nutrition and health [54, 55].

Since the pangenome of human *Limosilactobacillus* revealed high genomic heterogeneity (75.6% accessory genes), we analysed their accessory genome to understand whether there are signatures suggesting that these strains have specialized to succeed in the human body. Overall, the accessory genes of human *L. fermentum* do not cluster according to isolation sources, appearing to be ‘promiscuous’ in terms of human body sites. A free-living lifestyle was previously suggested for *L. fermentum*, both in terms of host range and body site [56]. In comparison, the accessory genes from urogenital *L. reuteri* clustered together, suggesting that the accessory genes were affected by the habitat. In fact, *L. reuteri* was previously suggested to diversify in a strict host-specific manner into host-adapted lineages by a long-term evolutionary process, allowing the development of a highly specialized symbiosis [57, 58]. Also, *L. mucosae* isolated from gastrointestinal tract or *L. coleohominis*, *L. vaginalis*, *L. portuensis*, *L. urinaemulieris* from urogenital tract were similar in the composition of their accessory genes, suggesting adaptation to specific human body sites. However, few genomes per species were available.

According to our enrichment analyses, KOs involved in fructoselysine degradation were detected only in some urogenital strains. Fructoselysine, abundant in cooked foods via the Maillard reaction, is a key product leading to the formation of glycation end products in the human body that have been associated with chronic diseases and development of diabetes complications and ageing. Previously, the conversion of fructoselysine has been reported for a few bacteria, including *Escherichia coli*, *Bacillus subtilis* and *Intestinimonas* [59]. Urogenital *Limosilactobacillus* might be able to catabolize fructoselysine from undigested food, as this Amadori product is excreted in urine, and the *frlD* and *frlB* genes encoding fructoselysine 6-kinase and fructoselysine 6-phosphate deglycase, respectively, were identified. However, it is not yet known whether urogenital *Limosilactobacillus* are able to grow on fructoselysine as the sole carbon and energy source.

AMR genes were found in higher frequency among strains isolated from animals, with few presenting multidrug-resistant genotypes. These findings may be explained by the extensive use of antibiotics in food-producing animals, since most of these strains were isolated from farm animals (e.g. chicken, pig) [60]. Moreover, the link between this antimicrobial usage in animals and the occurrence of AMR in the human microbiome/human infections has been deeply discussed in recent years [61, 62]. It should be noted, however, that most of AMR genes were identified in *L. reuteri*, which was also the species that showed less frequent CRISPR-Cas systems. This is in agreement with the idea that the consumption and dissemination of antibiotics in the environment is favouring the deficient forms of immunity provided by CRISPR-Cas systems [63].

Nearly one-third of *Limosilactobacillus* strains harboured the CRISPR-Cas system, being more frequent among strains recovered from humans, which might reflect the high prevalence of foreign DNA in the human host rather than in non-human sources, in particular the type II-A CRISPR-Cas system. The type II CRISPR-Cas system, a well-known molecular mechanism that provides adaptive immunity against exogenous genetic elements such as bacteriophages and plasmids in bacteria [64], might indicate an advantage in promoting *Limosilactobacillus* genome stability by acting as a barrier to entry of foreign DNA elements. However, the functionality of the identified type II-A CRISPR-Cas system must be investigated. As previously observed, most of the spacer sequences present in the CRISPR-Cas system remain without a match, representing the vast CRISPR 'dark matter' [65], which might be attributed to the presence of substantial mobile genetic elements that had not been sequenced to date [66].

A considerable diversity of bacteriocins, covering all classes, was found among *Limosilactobacillus* strains, which may contribute not only for a competitive advantage for the producing strain and modulate the neighbouring microbial community, but also benefit the host by inhibiting potential pathogens. Interestingly, our data suggests an enhanced ability of animal strains to produce different bacteriocins, with those from the animal encoding over 3.5 and 4.5 times as many putative bacteriocins as those from humans and food, respectively. Bacteriocin production might be a competitive advantage for strains from complex environments, such as the microbiota of animals. Enterolysin A, produced by some *Lactobacillus* species [67], was the most common bacteriocin, distributed among several niches (animal, food, human). This bacteriocin has been found to have a broad spectrum of activity due to its mode of action that results in cell wall degradation, being active against some pathogens (e.g. some strains of *Listeria monocytogenes* and *Staphylococcus aureus*), but with increased activity towards *Lactobacillaceae* [68, 69]. The second most common bacteriocin belonged to sactipeptides, also referred to as sactibiotics when they possess antimicrobial activity, and to date none from *Lactobacillaceae* have been characterized [70]. GassericinK7B, the third most common, was found among strains from animal origin, with the exception of one vaginal strain. It should be noted that gassericinK7B has shown potent activity against *Lactobacillus iners* strains that have been linked with the development of bacterial vaginosis and its production in strains from the urogenital tract may represent a beneficial trait [71]. Nevertheless, appropriate phenotypic tests should be performed to evaluate the expression of these genomic features.

Prophages are a common feature among prokaryotic genomes, including in *Lactobacillaceae* [72, 73]. In this study, approximately 40% of *Limosilactobacillus* human strains contained prophage regions, which is congruent with previous reports of the high prevalence of phages in strains colonizing humans, especially the gut [74]. Within the phage genes, we did not observe genes related to AMR or putatively related to bacterial virulence, which suggests that presence of bacteriophages in *Limosilactobacillus* strains can bring advantages on environmental adaptation rather than pathogenicity.

In summary, this study presents a comprehensive comparative genomic analyses of the genus *Limosilactobacillus* and provides a scientific basis toward understanding the biology and evolution of this genus. Here, we demonstrate high genomic diversity within *Limosilactobacillus*, the existence of putative novel species in the public databases and different bacterial lifestyles, depending on the species (free-living or more host-specific). Furthermore, a large number of accessory genes within *Limosilactobacillus* suggests their high species- and strain-specificity. According to the accessory genome, human *L. fermentum* appear to have a ubiquitous adaptation to different body sites, while for other species the evolution could be directed by a particular human body site (*L. reuteri*). Additionally, several common genomic determinants (e.g. putative bacteriocins or lactic acid production) support that *Limosilactobacillus* can have a role in maintaining host homeostasis. Nevertheless, further sequencing to enlarge the genome collection and appropriate phenotypic testing is essential to evaluate their potential contribution to human health.

Funding information

This work was financed by national funds from FCT – Fundação para a Ciência e a Tecnologia, I.P., in the scope of the project UIDP/04378/2020 and UIDB/04378/2020 of the Research Unit on Applied Molecular Biosciences – UCIBIO and the project LA/P/0140/2020 of the Associate Laboratory Institute for Health and Bioeconomy – i4HB. M.K. is supported by FCT (SFRH/BD/132497/2017). F.G. is supported by national funds through FCT in the context of the transitional norm (DL57/2016/CP1346/CT0034); and T.G.R. by UCIBIO – Applied Molecular Biosciences Unit (UIDP/QUI/04378/2020), with the financial support of the FCT/MCTES through national funds.

Acknowledgements
We would like to thank Svetlana Ugarcina Perovic and Joana Rocha for their contributions to the Female Urogenital Microbiome Project.

Author contributions

M.K., performed all genomic analyses, interpreted the data, performed visualizations and wrote the manuscript. F.G. and T.G.R., contributed to data interpretation, and revised and edited the manuscript. L.P., contributed to reviewing the manuscript, project design and administration, and funding.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

1. Ksiezarek M, Grosso F, Ribeiro TG, Peixe L. Genomic diversity of genus *Limosilactobacillus*. *Figshare*. 2022. DOI: 10.6084/m9.figshare.20052164.
2. Salvetti E, Harris HMB, Felis GE, O'Toole PW. Comparative genomics of the genus *Lactobacillus* reveals robust phylogroups that provide the basis for reclassification. *Appl Environ Microbiol* 2018;84:e00993-18.
3. Zheng J, Wittouck S, Salvetti E, Franz C, Harris HMB, *et al*. A taxonomic note on the genus *Lactobacillus*: description of 23 novel genera, emended description of the genus *Lactobacillus* Beijerinck 1901, and union of *Lactobacillaceae* and *Leuconostocaceae*. *Int J Syst Evol Microbiol* 2020;70:2782–2858.
4. Ksiezarek M, Ribeiro TG, Rocha J, Grosso F, Perovic SU, *et al*. *Limosilactobacillus urinaemulieris* sp. nov. and *limosilactobacillus portuensis* sp. nov. isolated from urine of healthy women. *Int J Syst Evol Microbiol* 2021;71:004726.
5. Li F, Cheng CC, Zheng J, Liu J, Quevedo RM, *et al*. *Limosilactobacillus balticus* sp. nov., *Limosilactobacillus agrestis* sp. nov., *Limosilactobacillus albertensis* sp. nov., *Limosilactobacillus rudii* sp. nov. and *Limosilactobacillus fastidiosus* sp. nov., five novel *Limosilactobacillus* species isolated from the vertebrate gastrointestinal tract, and proposal of six subspecies of *Limosilactobacillus reuteri* adapted to the gastrointestinal tract of specific vertebrate hosts. *Int J Syst Evol Microbiol* 2021;71:004644.
6. Naghmouchi K, Belguesmia Y, Bendali F, Spano G, Seal BS, *et al*. *Lactobacillus fermentum*: a bacterial species with potential for food preservation and biomedical applications. *Crit Rev Food Sci Nutr* 2020;60:3387–3399.
7. Mu Q, Tavella VJ, Luo XM. Role of *Lactobacillus reuteri* in human health and diseases. *Front Microbiol* 2018;9:757.
8. Pasolli E, De Filippis F, Mauriello IE, Cumbo F, Walsh AM, *et al*. Large-scale genome-wide analysis links lactic acid bacteria from food with the gut microbiome. *Nat Commun* 2020;11:2610.
9. Ksiezarek M, Ugarcina-Perovic S, Rocha J, Grosso F, Peixe L. Long-term stability of the urogenital microbiota of asymptomatic European women. *BMC Microbiol* 2021;21:64.
10. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
11. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;25:1043–1055.
12. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods* 2016;8:12–24.
13. Ramasamy D, Mishra AK, Lagier J-C, Padhmanabhan R, Rossi M, *et al*. A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *Int J Syst Evol Microbiol* 2014;64:384–391.
14. R Core Team. R: a Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing, 2018. https://www.R-project.org/
15. Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, *et al*. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* 2021;6:3–6.
16. Lee MD. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* 2019;35:4162–4164.
17. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 2009;26:1641–1650.
18. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;47:W256–W259.
19. Inkscape Project. Inkscape; 2020. https://inkscape.org
20. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, *et al*. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11:119.
21. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 2012;40:D109–D114.
22. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;32:1792–1797.
23. van Dongen S, Abreu-Goodger C. Using MCL to extract clusters from networks. *Methods Mol Biol* 2012;804:281–295.
24. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag, 2016. https://ggplot2.tidyverse.org
25. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, *et al*. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3693.
26. Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* 2011;12:35.
27. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 2013;8:e61217.
28. Shaiber A, Willis AD, Delmont TO, Roux S, Chen L-X, *et al*. Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol* 2020;21:292.
29. van Heel AJ, de Jong A, Song C, Viel JH, Kok J, *et al*. BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic Acids Res* 2018;46:W278–W281.
30. Russel J, Pinilla-Redondo R, Mayo-Muñoz D, Shah SA, Sørensen SJ. CRISPRCasTyper: automated identification, annotation, and classification of CRISPR-Cas loci. *CRISPR J* 2020;3:462–469.
31. Akhter S, Aziz RK, Edwards RA. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res* 2012;40:e126.
32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
33. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, *et al*. ResFinder 4.0 for predictions of phenotypes from genotypes. *J Antimicrob Chemother* 2020;75:3491–3500.
34. Cárdenas N, Laiño JE, Delgado S, Jiménez E, Juárez del Valle M, *et al*. Relationships between the genome and some phenotypical properties of *Lactobacillus fermentum* CECT 5716, a probiotic strain isolated from human milk. *Appl Microbiol Biotechnol* 2015;99:4343–4353.

35. Sun Z, Harris HMB, McCann A, Guo C, Argimón S, *et al*. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. *Nat Commun* 2015;6:8322.

36. Kant R, Blom J, Palva A, Siezen RJ, de Vos WM. Comparative genomics of *Lactobacillus*. *Microb Biotechnol* 2011;4:323–332.

37. Yu J, Zhao J, Song Y, Zhang J, Yu Z, *et al*. Comparative genomics of the herbivore gut symbiont *Lactobacillus reuteri* reveals genetic diversity and lifestyle adaptation. *Front Microbiol* 2018;9:1151.

38. Illeghems K, De Vuyst L, Weckx S. Comparative genome analysis of the candidate functional starter culture strains *Lactobacillus fermentum* 222 and *Lactobacillus plantarum* 80 for controlled cocoa bean fermentation processes. *BMC Genomics* 2015;16:766.

39. Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A, *et al*. The ecology and evolution of pangenomes. *Curr Biol* 2019;29:R1094–R1103.

40. Chen L, Gu Q, Li P, Chen S, Li Y. Genomic analysis of *Lactobacillus reuteri* WHH1689 reveals its probiotic properties and stress resistance. *Food Sci Nutr* 2019;7:844–857.

41. Xu S, Cheng J, Meng X, Xu Y, Mu Y. Complete genome and comparative genome analysis of *Lactobacillus reuteri* YSJL-12, a potential probiotics strain isolated from healthy sow fresh feces. *Evol Bioinform Online* 2020;16:1176934320942192.

42. Yoo D, Bagon BB, Valeriano VDV, Oh JK, Kim H, *et al*. Complete genome analysis of *Lactobacillus fermentum* SK152 from kimchi reveals genes associated with its antimicrobial activity. *FEMS Microbiol Lett* 2017;364:fnx185.

43. Brandt K, Nethery MA, O'Flaherty S, Barrangou R. Genomic characterization of *Lactobacillus fermentum* DSM 20052. *BMC Genomics* 2020;21:328.

44. Romero R, Hassan SS, Gajer P, Tarca AL, Fadrosh DW, *et al*. The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome* 2014;2:4.

45. Das Purkayastha S, Bhattacharya MK, Prasad HK, Upadhyaya H, Lala SD, *et al*. Contrasting diversity of vaginal lactobacilli among the females of Northeast India. *BMC Microbiol* 2019;19:198.

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

47. Belstrøm D, Constancias F, Liu Y, Yang L, Drautz-Moses DI, *et al*. Metagenomic and metatranscriptomic analysis of saliva reveals disease-associated microbiota in patients with periodontitis and dental caries. *NPJ Biofilms Microbiomes* 2017;3:23.

48. Million M, Maraninchi M, Henry M, Armougom F, Richet H, *et al*. Obesity-associated gut microbiota is enriched in *Lactobacillus reuteri* and depleted in *Bifidobacterium animalis* and *Methanobrevibacter smithii*. *Int J Obes* 2012;36:817–825.

49. Inglin RC, Meile L, Stevens MJA. Clustering of pan- and core-genome of *Lactobacillus* provides novel evolutionary insights for differentiation. *BMC Genomics* 2018;19:284.

50. Mendes-Soares H, Suzuki H, Hickey RJ, Forney LJ. Comparative functional genomics of *Lactobacillus* spp. reveals possible mechanisms for specialization of vaginal lactobacilli to their environment. *J Bacteriol* 2014;196:1458–1470.

51. Valeriano VDV, Oh JK, Bagon BB, Kim H, Kang D-K. Comparative genomic analysis of *Lactobacillus mucosae* LM1 identifies potential niche-specific genes and pathways for gastrointestinal adaptation. *Genomics* 2019;111:24–33.

52. Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, *et al*. Comparative genomics of the lactic acid bacteria. *Proc Natl Acad Sci USA* 2006;103:15611–15616.

53. Hernández-Montes G, Díaz-Mejía JJ, Pérez-Rueda E, Segovia L. The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution. *Genome Biol* 2008;9:R95.

54. Peterson CT, Rodionov DA, Osterman AL, Peterson SN. B vitamins and their role in immune regulation and cancer. *Nutrients* 2020;12:E3380.

55. Yoshii K, Hosomi K, Sawane K, Kunisawa J. Metabolism of dietary and microbial vitamin B family in the regulation of host immunity. *Front Nutr* 2019;6:48.

56. Verce M, De Vuyst L, Weckx S. Comparative genomics of *Lactobacillus fermentum* suggests a free-living lifestyle of this lactic acid bacterial species. *Food Microbiol* 2020;89:103448.

57. Oh PL, Benson AK, Peterson DA, Patil PB, Moriyama EN, *et al*. Diversification of the gut symbiont *Lactobacillus reuteri* as a result of host-driven evolution. *ISME J* 2010;4:377–387.

58. Frese SA, Benson AK, Tannock GW, Loach DM, Kim J, *et al*. The evolution of host specialization in the vertebrate gut symbiont *Lactobacillus reuteri*. *PLoS Genet* 2011;7:e1001314.

59. Bui TPN, Ritari J, Boeren S, de Waard P, Plugge CM, *et al*. Production of butyrate from lysine and the Amadori product fructoselysine by a human gut commensal. *Nat Commun* 2015;6:10062.

60. Pokharel S, Shrestha P, Adhikari B. Antimicrobial use in food animals and human health: time to implement "One Health" approach. *Antimicrob Resist Infect Control* 2020;9:181.

61. Lazarus B, Paterson DL, Mollinger JL, Rogers BA. Do human extraintestinal *Escherichia coli* infections resistant to expanded-spectrum cephalosporins originate from food-producing animals? A systematic review. *Clin Infect Dis* 2015;60:439–452.

62. Aarestrup FM. The livestock reservoir for antimicrobial resistance: a personal view on changing patterns of risks, effects of interventions and the way forward. *Philos Trans R Soc Lond B Biol Sci* 2015;370:20140085.

63. Butiuc-Keul A, Farkas A, Carpa R, Iordache D. CRISPR-Cas system: the powerful modulator of accessory genomes in prokaryotes. *Microb Physiol* 2022;32:2–17.

64. Chylinski K, Makarova KS, Charpentier E, Koonin EV. Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Res* 2014;42:6091–6105.

65. Shmakov SA, Sitnik V, Makarova KS, Wolf YI, Severinov KV, *et al*. The CRISPR spacer space is dominated by sequences from species-specific mobilomes. *mBio* 2017;8:e01397-17.

66. Yang L, Li W, Ujiroghene OJ, Yang Y, Lu J, *et al*. Occurrence and diversity of CRISPR loci in *Lactobacillus casei* group. *Front Microbiol* 2020;11:624.

67. de Jesus LCL, Drumond MM, Aburjaile FF, Sousa T de J, Coelho-Rocha ND, *et al*. Probiogenomics of *Lactobacillus delbrueckii* subsp. *lactis* cidca 133: in silico, in vitro, and in vivo approaches. *Microorganisms* 2021;9:829.

68. Khan H, Flint SH, Yu P-L. Determination of the mode of action of enterolysin A, produced by *Enterococcus faecalis* B9510. *J Appl Microbiol* 2013;115:484–494.

69. Nilsen T, Nes IF, Holo H. Enterolysin A, A cell wall-degrading bacteriocin from *Enterococcus faecalis* LMG 2333. *Appl Environ Microbiol* 2003;69:2975–2984.

70. Alvarez-Sieiro P, Montalbán-López M, Mu D, Kuipers OP. Bacteriocins of lactic acid bacteria: extending the family. *Appl Microbiol Biotechnol* 2016;100:2939–2951.

71. Nilsen T, Swedek I, Lagenaur LA, Parks TP. Novel selective inhibition of *Lactobacillus iners* by *Lactobacillus*-derived bacteriocins. *Appl Environ Microbiol* 2020;86:e01594-20.

72. Davidson BE, Kordias N, Dobos M, Hillier AJ. Genomic organization of lactic acid bacteria. *Antonie Van Leeuwenhoek* 1996;70:161–183.

73. Mercanti DJ, Carminati D, Reinheimer JA, Quiberoni A. Widely distributed lysogeny in probiotic lactobacilli represents a potentially high risk for the fermentative dairy industry. *Int J Food Microbiol* 2011;144:503–510.

74. Manrique P, Dills M, Young MJ. The human gut phage community and its implications for health and disease. *Viruses* 2017;9:E141.