ARTICLE
Cancer Advances 2023;6:e23018. https://doi.org/10.53388/2023623018

Comparative genomic analysis of *Lactobacillus crispatus* strains Lc31 and Lc83 with anti-cervical cancer potential by complete genome sequencing

Xue-Qin Cai[1#], Ruo-Nan Li[2#], Yan Ma[3], Xi-Xi Chen[3], Li-Juan Wan[1], Jin-Song Kan[1], Hui-Yan Wang[1*]

[1]Department of Laboratory Diagnostics, the First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230001, China. [2]Department of Obstetrics and Gynecology, the First Affiliated Hospital of Chongqing Medical University, Chongqing 400016, China. [3]Department of Gynecologic Oncology, the First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230001, China.

[#]Xue-Qin Cai and Ruo-Nan Li are the co-first authors of this paper.

***Corresponding to:** Hui-Yan Wang, Department of Laboratory Diagnostics, the First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, 107 Huan Hu Dong Lu, Shushan District, Hefei 230001, China. E-mail: wanghuiyanyx@ustc.edu.cn.

Author contributions

WH designed the study and supervised manuscript preparation. CX performed the analyses except MGE analysis. LR prepared the draft manuscript. MY prepared all figures and tables. CX collected and characterized the bacterial strains. WL performed the carbohydrate fermentation experiments. KJ performed the MGE analysis. All the authors have read and approved the final version of the manuscript.

Competing interests

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by grants from the Natural Science Foundation of Anhui Province (grant number 2208085MH253); the National Natural Science Foundation (grant number 81702560); and the Fundamental Research Funds for the Central Universities (grant number WK9110000110), People's Republic of China.

Peer review information

Cancer Advances thanks all Xiu-Shan Feng, Xiao-Tong Yao and Fang Yang for their contribution to the peer review of this paper.

Abbreviations

UGT, urogenital tract; GIT, gastrointestinal tract; IS, insertion sequence; STIs, sexually transmitted infections; hrHPV, high-risk HPV; CIN, cervical intraepithelial neoplasia; MRS, deMan Rogosa Sharpe; SEM, Scanning electron microscopy; CDS, protein-coding sequences; BLAST, Basic Local Alignment Search Tool; KEGG, Kyoto Encyclopedia of Genes and Genomes; COG, Clusters of Orthologous Groups; NR, NCBI nonredundant protein; GO, Gene Ontology; ML, maximum-likelihood; MP, maximum parsimony; MLSA, multilocus sequence analysis; TYGS, Type (Strain) Genome Server; ANI, Average nucleotide identity; CV, Composition Vector; GBDP, genome blast distance phylogeny method; NJ, neighbor-joining; WGA, Whole-genome alignment; LCB, locally collinear block; MGE, Mobile genetic element; GIs, Genomic islands; BGCs, biosynthetic gene clusters; VF, virulence factor; RGI, Resistance Gene Identifier; CARD, Comprehensive Antibiotic Resistance Database; SD, standard deviation; CLSI, Clinical and Laboratory Standards Institute; HGT, horizontal gene transfer; LLS, listeriolysin S; LAB, lactic acid bacteria; GHs, glycoside hydrolases; GTs, glycosyl transferases; CBMs, noncatalytic carbohydrate-binding modules; CEs, carbohydrate esterases; AA, auxiliary activity.

Citation

Cai XQ, Li RN, Ma Y, et al. Comparative genomic analysis of *Lactobacillus crispatus* strains Lc31 and Lc83 with anti-cervical cancer potential by complete genome sequencing. *Cancer Adv.* 2023;6:e23018. doi: 10.53388/2023623018.

Executive editor: Zi-Yao Feng.

Received: 13 July 2023; **Accepted:** 05 September 2023; **Available online:** 26 September 2023.

Abstract

Lactobacillus crispatus is a commonly found species in the urogenital tract (UGT) of healthy females and can also colonize other niches, such as the gastrointestinal tract (GIT). Although its potential protective role in cervical cancer has been reported, the anticancer mechanisms involved are still unclear. In this study, we sequenced and characterized the complete genomes of two *L. crispatus* strains (Lc31 and Lc83) isolated from the UGT of healthy women of reproductive age. Phylogenetic and phylogenomic analyses of these two strains and 15 other *L. crispatus* strains with complete genome sequences revealed that strains from the UGT and GIT clustered separately. UGT strains had a larger genome size, higher GC contents, and more protein-coding sequences and insertion sequence (IS) elements, indicating the likelihood of active horizontal gene transfer in this niche. We found a universal presence of genes encoding bacteriocins and the absence of virulence factors and antibiotic resistance genes in UGT strains, suggesting the potential of *L. crispatus* as a urogenital probiotic. Comparative genomic analysis identified an *ula* gene cluster responsible for L-ascorbate catabolism exclusively in UGT strains, and carbohydrate fermentation experiments confirmed that this substrate supported the growth of *L. crispatus* Lc31 and Lc83. Our findings improve the understanding of how the genome determines niche adaptation by *L. crispatus*, providing a foundation for investigating the mechanisms by which urogenital-derived *L. crispatus* promotes female health.

Keywords: *Lactobacillus crispatus*, anti-cervical cancer, genetic diversity, comparative genomics, antibiotic susceptibility, L-ascorbate catabolism

Introduction

In the lower genital tract of healthy women, lactobacilli are the predominant microorganisms that benefit the host by protecting the vaginal microenvironment against invading pathogens. Depletion of lactobacilli increases the risk of sexually transmitted infections (STIs), premature birth, and other obstetric and gynecologic problems, including cancers [1]. The genus *Lactobacillus* comprises over 200 species, but only four species (*Lactobacillus crispatus*, *L. gasseri*, *L. iners*, and *L. jensenii*) are common in the vaginal microbiota of healthy women of reproductive age [2]. The prevalence and abundance of the four species vary considerably: *L. crispatus* and *L. iners* are more frequently detected than *L. gasseri* and *L. jensenii* [3]. Furthermore, *L. iners* is detected in both healthy vaginas and during vaginal dysbiosis, while *L. crispatus* tends to be detected only in healthy vaginas [4]. Vaginal microbiomes that are dominated by *L. iners* are less stable, as they frequently transform to non–*Lactobacillus*-dominated communities [1, 4]. In contrast, a predominance of *L. crispatus* is considered optimal for vaginal health, as it is generally associated with a lower vaginal pH and a reduced risk of STIs [1, 2].

Persistent high-risk HPV (hrHPV) infection is the leading cause of cervical intraepithelial neoplasia (CIN) and cancer [5]. As shown in our previous study, *L. crispatus* is significantly negatively correlated with hrHPV infection and CIN and plays a potential protective role in the occurrence and development of cancer, which is expected to become a new technical means to prevent and treat cervical cancer [3]. Although this phenomenon has been confirmed by multiple studies, the mechanisms involved are still unclear [6, 7]. Interestingly, in addition to the lower genital tract, human urinary and gastrointestinal tracts are also common sites for *L. crispatus* colonization [8, 9]. In addition, *L. crispatus* can also be isolated from the gastrointestinal tract (GIT) of poultry and animals [10, 11]. Whether there are differences between these strains from different isolation sources is crucial for the application of *L. crispatus* as an anticancer probiotic. Therefore, it is necessary to clarify the intraspecies diversity of *L. crispatus* and reveal the underlying genetic mechanisms.

Genome sequencing and comparative genomics are powerful tools for studying bacterial diversity and have been applied in various probiotic studies [12–14]. In this study, two *L. crispatus* strains, Lc31 and Lc83, were isolated from the vaginas of healthy women of reproductive age. The complete genomes of these two strains were sequenced and compared with those of 15 other strains of *L. crispatus* with complete genome sequences. Phylogenetic and phylogenomic approaches were used to explore the intraspecies diversity among the strains of *L. crispatus*. Genome mining of genetic elements and comparative genomics were conducted to identify core, pan, and specific genes, revealing the genetic basis of diversity among *L. crispatus* strains. Additionally, carbohydrate fermentation experiments and antibiotic susceptibility tests were performed to validate the results of the comparative genomic analyses. The findings from this study can provide new insights into the genetic determinants of probiotic characteristics in *L. crispatus*.

Materials and methods

Strain isolation and identification

Sterile swabs were used to collect vaginal secretions from two healthy women of reproductive age, which were then eluted with 1 mL physiological saline and cultured overnight at 37 ˚C in 10 mL deMan Rogosa Sharpe (MRS) broth. Informed written consent was obtained from both participants. After enrichment, cultures were plated on MRS solid medium, and two bacterial strains were isolated by repeated plate streaking, Gram staining, and light microscopy examination using an Olympus BX53 microscope. Cell pellets of both strains were collected by centrifugation, washed three times with ice-cold phosphate-buffered saline, and resuspended in 2.5% glutaraldehyde. Scanning electron microscopy (SEM) was performed to examine cell surface morphology.

Genomic DNA was extracted from the isolated strains using an E.Z.N.A. bacterial DNA Kit (D3350-01, Omega Biotek, Norcross, GA, USA) and purified. The 16S rRNA gene sequence was amplified by PCR with universal bacterial primers 27F (5'-AGAGTTTGATCCTGGCTCAG-3') and 1492R (5'-GGTTACCTTGTTACGACTT-3'). The PCR products were then sequenced by Beijing Genomics Institute (BGI; Shanghai, China). The study received ethics approval from the Institutional Review Board of the First Affiliated Hospital of USTC (no.2020-JYK-01). The isolated strains were deposited in the China General Microbiological Culture Collection Centre under accession numbers CGMCC 21348 for strain Lc31 and CGMCC 21349 for strain Lc83.

Genome sequencing and assembly

The two strains were cultured overnight, and the resulting cell pellets were collected and sent on dry ice to BGI (Wuhan, China) for library construction and sequencing. Genomic DNA was isolated from each of the cell pellets using the DNeasy Blood and Tissue Kit (69504, Qiagen, Hilden, Germany), and DNA quality was examined by agarose gel electrophoresis and a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific). To construct the 10-kb DNA libraries, a PacBio SMRTbell 10 kb Library Preparation Kit (Pacific Biosciences, Menlo Park, CA, USA) was used, while paired-end libraries were constructed with a NEBNext Ultra DNA Library Prep Kit (E7370L, Illumina, San Diego, CA, USA) according to the manufacturer's instructions. Both PacBio RS II (Pacific Biosciences) and HiSeq 4000 (Illumina) platforms were used for genome sequencing. Draft genomic unitigs were assembled using Celera Assembler v8.3 against a set of subreads with a high-quality, corrected, circular, consensus sequence [15]. Single-base corrections were made with Illumina short reads (PE150) using the GATK and SOAP tool packages (SOAP2, SOAPsnp, and SOAPindel) to improve the accuracy of genome sequences [16, 17]. The Illumina reads were filtered and mapped to the bacterial plasmid database (http://www.ebi.ac.uk/genomes/plasmid.html) using SOAP to detect the presence of any plasmid. PacBio long reads and Illumina short reads were both used for plasmid sequence assembly in the same way as for genome sequence assembly.

Genome component prediction and gene annotation

Gene predictions for the assembled genomes were generated using Glimmer v3.02 with hidden Markov models [18]. tRNAs and rRNAs were identified using tRNAscan-SE [19] and RNAmmer [20], respectively. For functional annotation, protein-coding sequences (CDS) were predicted and categorized using Basic Local Alignment Search Tool (BLAST) against seven databases including Kyoto Encyclopedia of Genes and Genomes (KEGG) [21], Clusters of Orthologous Groups (COG) [22], NCBI nonredundant protein (NR), Gene Ontology (GO) [23], Nonsupervised Orthologous Groups (EggNOG) [24], Swiss Institute of Bioinformatics protein (Swiss-Prot), and Translated European Molecular Biology Laboratory nucleotide (TrEMBL) databases [25]. The web-based service of ContEst16S was used to detect possible biological contamination in genome assemblies [26].

Taxonogenomic assessments

To elucidate the taxonomy and phylogenetic relationship of the two strains of *Lactobacillus crispatus*, the complete genome sequences of the 15 other strains of *L. crispatus* were retrieved from the NCBI database, and 16S rRNA gene sequences between primers 27F and 1492R were extracted from all strains and aligned pairwise using the nucleotide BLAST algorithm. However, due to the sequence conservation of the 16S rRNA gene, it has limited utility for distinguishing between closely related strains within a species [27, 28]. To overcome this limitation, alternative phylogenetic markers and multiple phylogenomic approaches were adopted. The *recA* gene provides a detailed phylogenetic resolution of relationships among species of the genus *Lactobacillus* [28, 29]. Therefore, *recA* gene sequences were also extracted from the 17 *L. crispatus* genomes. Multiple sequence

alignment was performed using Muscle [30], and phylogenetic trees (cladograms) were reconstructed using maximum-likelihood (ML) and maximum parsimony (MP) methods employed in the MEGA X software package [31]. The statistical reliability of the trees was evaluated by bootstrap analysis with 1000 replicates.

To reduce possible biases in the phylogenetic analyses caused by the dependence on using only one gene, the phylogenetic relationships among the strains of *L. crispatus* were also analyzed by multilocus sequence analysis (MLSA) using the protein products of 31 bacterial housekeeping genes [32]. Briefly, the genomes of the 17 *L. crispatus* strains were submitted to the webserver AmphoraNet to extract the 31 universally conserved proteins [32]. The protein sequences of the individual marker genes were concatenated for individual genomes and then aligned using Muscle. Following the alignment, cladograms were reconstructed in MEGA X software using ML and MP methods with 1000 bootstrap replicates.

Phylogenomic assessments

A set of well-established phylogenomic approaches were employed to assess the phylogeny with whole genome data. Digital DNA–DNA hybridization (dDDH) was performed using the Type (Strain) Genome Server (TYGS) [33]. Average nucleotide identity (ANI) values were calculated from whole genome sequences using the JSpecies software program and visualized as a heatmap using Morpheus software [34]. Additionally, a whole-genome-based tree without sequence alignment was also reconstructed with Composition Vector Tree v3 (CVTree3) and a K value of 6 using the protein sequences of *L. crispatus* strains [35]. A genome blast distance phylogeny method (GBDP) tree was also created using the TYGS platform [33]. The core gene alignment from Roary was used in TreeBeST with 1000 bootstrap iterations to reconstruct a neighbor-joining (NJ) phylogenetic tree. The phylogenetic trees were then visualized and annotated using iTOL (https://itol.embl.de/) [36].

Whole-genome alignment (WGA)

Multiple and pairwise whole-genome alignments of the 17 *L. crispatus* genomes were conducted using the Mauve aligner (version 2015-02-25) with the progressive Mauve algorithm and default scoring and parameters [37]. The generated locally collinear block (LCB) order and orientation were input into the GRIMM web server (http://grimm.ucsd.edu/MGR/) to calculate the pairwise genome rearrangement distance between the type strain ATCC 33820 and the other strains [38]. To further dissect the collinearity, pairwise whole-genome dot-plots were created using YASS with the following settings: E value of 1.0×10^{-30}; X-drop of 50; window range of 100–200000; window increment of 2 ×; the hit criterion of double, and default settings for others [39].

Mobile genetic element (MGE) analysis

Prophage regions were identified with PHAge Search Tool Enhanced Release (PHASTER) [40], but only intact regions were included. CRISPR loci and *cas* genes were searched using CRISPRCasFinder with the default parameters [41]. Only the CRISPR loci accompanied by *cas* genes in their vicinity were regarded as confirmed CRISPRs and included. Genomic islands (GIs) were predicted using IslandViewer 4, which integrates IslandPick, SIGI-HMM, and IslandPath-DIMO [42]. Insertion sequences (ISs) were detected by the web tool MobileElementFinder version 1.0.3 with the parameters of 95% minimum alignment coverage, 90% minimum sequence identity, and 30 nt maximum truncation [43].

Genome mining of secondary metabolite biosynthetic gene clusters (BGCs), virulence factors (VFs) and antibiotic resistance genes

Secondary metabolite biosynthetic gene clusters (BGCs) in the *L. crispatus* genomes were identified using the bacterial version of antiSMASH v6.0.1 [44], setting the detection strictness as relaxed and the extra features as all on. Putative virulence factor (VF) genes in the genomes were assessed by the BLASTP program of the VFanalyzer pipeline against the selected database of "VFDB core dataset - proteins associated with experimentally verified VFs" with a threshold of 50% pairwise amino acid identity and 0.7 length difference cutoff [45]. Antibiotic resistance genes were searched using the Resistance Gene Identifier (RGI) tool of the Comprehensive Antibiotic Resistance Database (CARD), processing the protein sequence files with the criteria "perfect, strict, complete genes only" [46].

Comparative genomic analysis

For comparative genomic analysis, all 17 *L. crispatus* genomes underwent core- and pangenome predictions. Genes were categorized based on their distribution across the genomes. Shared genes present in all strains were considered core genes, while genes present in some of the strains were considered dispensable genes. Genes assigned to a single strain were considered unique. CD-HIT v4.6.6 software was used for rapid clustering of similar proteins using a threshold of 50% pairwise amino acid identity and a 0.7 length difference cutoff to compute the core genes [47]. The pangenome consisted of all three types of genes, including the niche-specific genes that were filtered from the dispensable genes. All gene sets were annotated with COG categories using eggnog-mapper v2 based on eggNOG v5.0 with the default settings [24]. Additionally, the KEGG Automatic Annotation Server Ver.2.1 was used to predict metabolic pathways in the individual genomes [21].

Carbohydrate-active enzyme (CAZyme) analysis

Identification of CAZymes across the 17 *L. crispatus* genomes was determined using the dbCAN2 meta server with three annotation tools, HMMER, DIAMOND, and Hotpep [48]. The candidates found by at least two tools were included in the analysis of CAZyme classes and families.

Carbohydrate fermentation

Strains were cultured in 25 mL test tubes containing 10 mL of modified Morishita minimal medium, adjusted to pH 6.4 and sterilized by filtration at 0.2 μm [49]. Solutions of carbon sources, including glucose, glycogen, and L-ascorbate, were sterilized by filtration at 0.2 μm and added to the medium to a final concentration of 10 g/L. Cell pellets of *L. crispatus* Lc31 and Lc83 were collected from overnight MRS cultures, washed three times with minimal medium, and then added to the test tube to achieve a final concentration of 5×10^6 CFU/mL. Growth was monitored by means of McFarland turbidimetry with a DensiCHEK Plus device (bioMérieux, France). All experiments were performed in triplicate, and the mean ± standard deviation (SD) was calculated.

Antibiotic susceptibility assay

The antibiotic susceptibility of *L. crispatus* Lc31 and Lc83 was determined with the disc diffusion method as described previously [50]. Briefly, an overnight culture of *L. crispatus* was adjusted to 0.5 McFarland standard, and then 100 μL of this culture was spread on MRS solid plates. After the surface dried, the antibiotic discs were applied to the surface of the plates with forceps. The plates were then incubated at 37 ˚C for 24 h. A total of 18 antibiotics were separately used to detect the antibiotic susceptibility of *L. crispatus* Lc31 and Lc83. The results were interpreted as resistant (R), intermediate (I) or susceptible (S) according to the recommended standards based on the Clinical and Laboratory Standards Institute (CLSI) [50].

Statistical analysis

Statistical analyses were conducted using GraphPad Prism 8 (GraphPad Software, San Diego, CA, USA). The Mann–Whitney test was employed to compare the two groups, and a p value less than 0.05 was deemed statistically significant. Quantitative data are presented as the mean ± SD.

Data availability

The genome sequences for *L. crispatus* strains Lc31 and Lc83 have been deposited in the NCBI database under BioSample numbers

SAMN15862579 and SAMN15862762, respectively. The sequences of chromosomes are available via GenBank under accession numbers CP061006 and CP061005. The plasmid sequences are available with accession numbers ON569381 and ON569382, respectively.

Results

General genomic features

Cells of both obtained strains were gram-positive, and SEM revealed a straight or slightly curved rod-shaped morphology (Figure S1). The 16S rRNA gene sequences of both strains were > 99.9% identical to that of the type strain of the species, *L. crispatus* ATCC 33820 (Table S1). The observation that the 16S rRNA gene sequence similarities of the two strains with the type strain ATCC 33820 were greater than the 99% similarity threshold suggested that both strains belonged to the species *L. crispatus*. Consequently, the strains were designated *L. crispatus* Lc31 and Lc83, respectively. Sequencing and *de novo* assembly of the genome of strain Lc31 showed that it consisted of a 2,321,025-bp chromosome and a 16,664-bp plasmid, with 2319 CDS, 5 rRNA loci, and 72 tRNAs. The genome of strain Lc83 comprised a 2,308,435-bp chromosome and a 180,145-bp plasmid, with 2438 predicted CDS, 5 rRNA loci, and 72 tRNAs (Table 1). Genome maps (including both chromosomes and plasmids) of *L. crispatus* Lc31 and Lc83 are shown in Figure 1.

Although hundreds of *L. crispatus* genomes are present in the NCBI genome database, most are incomplete at the contig or scaffold level. Complete genome sequences for intraspecies comparison of genomic features were only available for 15 strains of *L. crispatus* in the NCBI Genome database. These genomes were retrieved and then screened for any possible biological contamination using the web server ContEst16S. The genomes of *L. crispatus* Lc31 and Lc83 obtained in this study were also screened with ContEst16S. By comparing five 16S rRNA gene fragments, none of the 17 genomes in this study appeared to be contaminated. The general features of the 17 complete genomes are listed in Table 1. The average chromosome size was 2.30 ± 0.15 Mb, and the GC content was 37.10 ± 0.12%. Plasmids were detected in 10 of the 17 strains of *L. crispatus*, with plasmid sizes ranging from 9 to 200 kb. An average of 2296 CDSs were predicted in the genomes. The number of tRNA genes varied between 64 and 73 (average 68). Most strains included 5 rRNA loci in their genome except strain ST1, which contained only 4 rRNA loci.

Taxonogenomics

The phylogeny within the species *L. crispatus* was described using the traditional 16S rRNA gene, the *recA* gene, and 31 bacterial single-copy genes. To strengthen the phylogenetic results, phylogenomic approaches based on whole-genome sequences were also employed, including digital DNA–DNA hybridization (dDDH), average nucleotide identity (ANI), composition vector (CV), and genome blast distance phylogeny (GBDP) analysis.

The 16S rRNA gene sequences of the 17 strains of *L. crispatus* showed > 99% identity in all pairwise comparisons (Table S1), indicating that all 17 strains were the same species. This classification was supported by pairwise dDDH calculations (Table S2). The highest dDDH value (99.1%) was between strains Lc31 and PMC201, while the lowest dDDH value (71.9%) was between strains Lc83 and C25. All dDDH values were greater than the recommended cutoff of 70% for prokaryotic species delineation [51], suggesting that these 17 strains belonged to the species *L. crispatus*. In addition, ANI values ranging from 96.15 to 100% were obtained with all these strains when compared with the type strain *L. crispatus* ATCC 33820 (Table S3), indicating that all these strains can be assigned to the same species of *L. crispatus* [34].

The *recA* gene is a more reliable marker for inferring phylogenetic relationships between closely related strains [28, 29]. The maximum likelihood (ML) tree based on *recA* gene sequences gave rise to two distinct clades (Figure 2A). Clade 1 comprised 11 strains of *L. crispatus* isolated from human urogenital samples (including vaginal swabs and urine samples), while clade 2 included the remaining six strains of *L. crispatus*, which were from gastrointestinal samples of humans, animals and poultry (including oral cavity, gut, and fecal samples). Trees generated by the maximum parsimony (MP) method had the same topology (Figure S2), and multilocus sequence analysis (MLSA) trees (Figure 2B and Figure S3) were also congruent with the niche-dependent clustering shown by the phylogenetic trees based on the *recA* gene.

Phylogenomics

Phylogenetic analyses based on the ANI, CV, and GBDP analyses (Figure 2C–E), consistently supported the results obtained from the *recA* phylogenetic analysis. These three whole-genome sequence-based approaches provide sufficient resolution at the species level and below [33–35]. Although the algorithms used in these analyses differ, the overall architecture of the resulting phylogenetic trees was consistent, supporting the reliability of the results. This was in agreement with the phylogenetic tree that was reconstructed based on the core genes of these 17 *L. crispatus* genomes (Figure 2F). Based on the phylogenetic and phylogenomic results, urogenital and gastrointestinal strains clustered separately in these trees, indicating that niche-specific adaptation may have played an important role in the evolutionary history of *L. crispatus*. To explore the underlying genetic mechanisms of niche-specific adaptation, the strains of *L. crispatus* in this study were divided into two groups: the urogenital group (clade 1) and the gastrointestinal group (clade 2). The gastrointestinal group contains strains from humans, chickens, and horses, while urogenital strains are exclusively from humans.

WGA

Multiple whole-genome alignment using ProgressiveMauve revealed that the 17 genomes of *L. crispatus* shared similarities in genomic content but not in synteny (Figure S4). A total of 53 LCBs were identified by Mauve, which represented regions of conservation and were characteristic of closely related genomes. However, numerous genome rearrangements were also detected by Mauve (Figure S4). Using the *L. crispatus* type strain ATCC 33820 as the reference, genome rearrangements were observed in most of the strains, except for the PRL2021 and Lc1226 strains (Table S4). A rearrangement distance of 9 was identified between the genomes of *L. crispatus* FDAARGOS_743 and ATCC 33820, which can be considered a hypothetical nine-step rearrangement process to transform the genome topology of one strain to the other. The genome rearrangements were evident in the pairwise alignments generated by YASS (Figure S5). Detailed sequence examination showed that these rearrangement regions were commonly associated with the presence of mobile genetic elements (MGEs), particularly ISs.

MGEs

The genomes of all 17 strains of *L. crispatus*, including Lc31 and Lc83, contained numerous MGEs, including prophages, CRISPRs, GIs, and ISs (Tables S5–S8). Intact prophages were identified in 13 of the 17 strains of *L. crispatus*; the exceptions were strains FDAARGOS_743, ATCC 33820, ST1, and C25. Of the 13 strains carrying intact prophages, strains Lc116, Lc1226, Lc1700, and 1D harbored two intact prophages, while the other nine strains each harbored a single intact prophage (Table S5). Phiadh was the most prevalent *Lactobacillus* phage in the current study and was detected in eight of the 11 urogenital strains (Table S5). Phiadh is a temperate phage that was first identified in *L. gasseri*; its presence in the human oropharynx was correlated with the prevalence of immunological disorders, although the underlying mechanism is not yet known [52]. The origin and function of *Lactobacillus* phage phiadh in the human urogenital tract also remains to be elucidated.

CRISPR is a well-known microbial defense mechanism against invading bacteriophages and plasmids. All 17 strains of *L. crispatus* harbored 2–6 CRISPR arrays, but not all harbored the *cas* gene locus genital strains (Table S6). The presence of the type I-E *cas* gene cluster in the vicinity of CRISPR arrays in strains Lc83, ATCC 33820, B4, ST1, and C25 indicated that they were type I-E CRISPR systems [53]. *L.*

Cancer Advances 2023;6:e23018. https://doi.org/10.53388/2023623018

Table 1 General features of the complete genomes of the 17 strains of *L. crispatus* included in this study

Strain	Biosample No.	GenBank accession	Source of isolation	Chromosome size	GC %	Plasmid (size)	Predicted CDS	tRNA	rRNA locus	GI	IS
Lc31	SAMN15862579	GCA_021278945.1	Human vagina	2,321,025 bp	37.17	plasmid 1 (16,664 bp)	2319	72	5	21	16
Lc83	SAMN15862762	GCA_021278925.1	Human vagina	2,308,435 bp	37.00	plasmid 1 (180,145 bp)	2438	68	5	20	16
AB70	SAMN08409124	GCA_003971565.1	Human vagina	2,351,263 bp	37.28	plasmid 1 (16,662 bp)	2331	72	5	29	14
FDAARGOS_743	SAMN11056458	GCA_009730275.1	Human vagina	2,373,487 bp	37.13	plasmid 1 (13,473 bp)	2325	65	5	19	43
CO3MRSI1	SAMN10343598	GCA_003795065.1	Human vagina	2,345,902 bp	37.09	n.d.	2242	66	5	24	25
PRL2021	SAMN15357510	GCA_016767795.1	Human vagina	2,411,492 bp	37.31	n.d.	2320	68	5	22	17
2029	SAMN02298797	GCA_000466885.3	Human vagina	2,500,140 bp	37.10	plasmid 1 (16,663 bp)	2496	70	5	20	25
PMC201	SAMN19601668	GCA_018885325.1	Human vagina	2,315,587 bp	37.24	plasmid 1 (25,315 bp); plasmid 2 (41,535 bp); plasmid 3 (9,163 bp)	2373	71	5	15	19
ATCC 33820	SAMN18472633	GCA_018987235.1	Human oral cavity	2,239,089 bp	37.00	n.d.	2112	64	5	20	30
Lc116	SAMN21367957	GCA_020042225.1	Human urine	2,350,925 bp	37.15	plasmid 1 (96,317 bp)	2393	70	5	21	15
Lc1226	SAMN21367958	GCA_020042125.1	Human urine	2,526,154 bp	37.12	n.d.	2507	72	5	23	14
Lc1700	SAMN21367959	GCA_020042005.1	Human urine	2,423,824 bp	37.24	plasmid 1 (200,485 bp); plasmid 2 (194,655 bp)	2743	70	5	23	20
B4	SAMN15516019	GCA_013456995.1	Human gut	2,039,590 bp	37.02	n.d.	1965	65	5	17	6
ST1	SAMEA2272191	GCA_000091765.1	Chicken crop	2,043,161 bp	36.89	n.d.	2024	64	4	22	6
C25	SAMN13634618	GCA_009933525.1	Chicken cecum	2,145,933 bp	37.01	plasmid 1 (179,151 bp)	2218	66	5	15	8
DC21.1	SAMN11372136	GCA_009769205.1	Chicken cecum	2,023,061 bp	37.03	plasmid 1 (16,343 bp)	1937	64	5	10	6
1D	SAMN13624824	GCA_013487905.1	Horse feces	2,349,358 bp	36.93	n.d.	2291	73	5	23	7

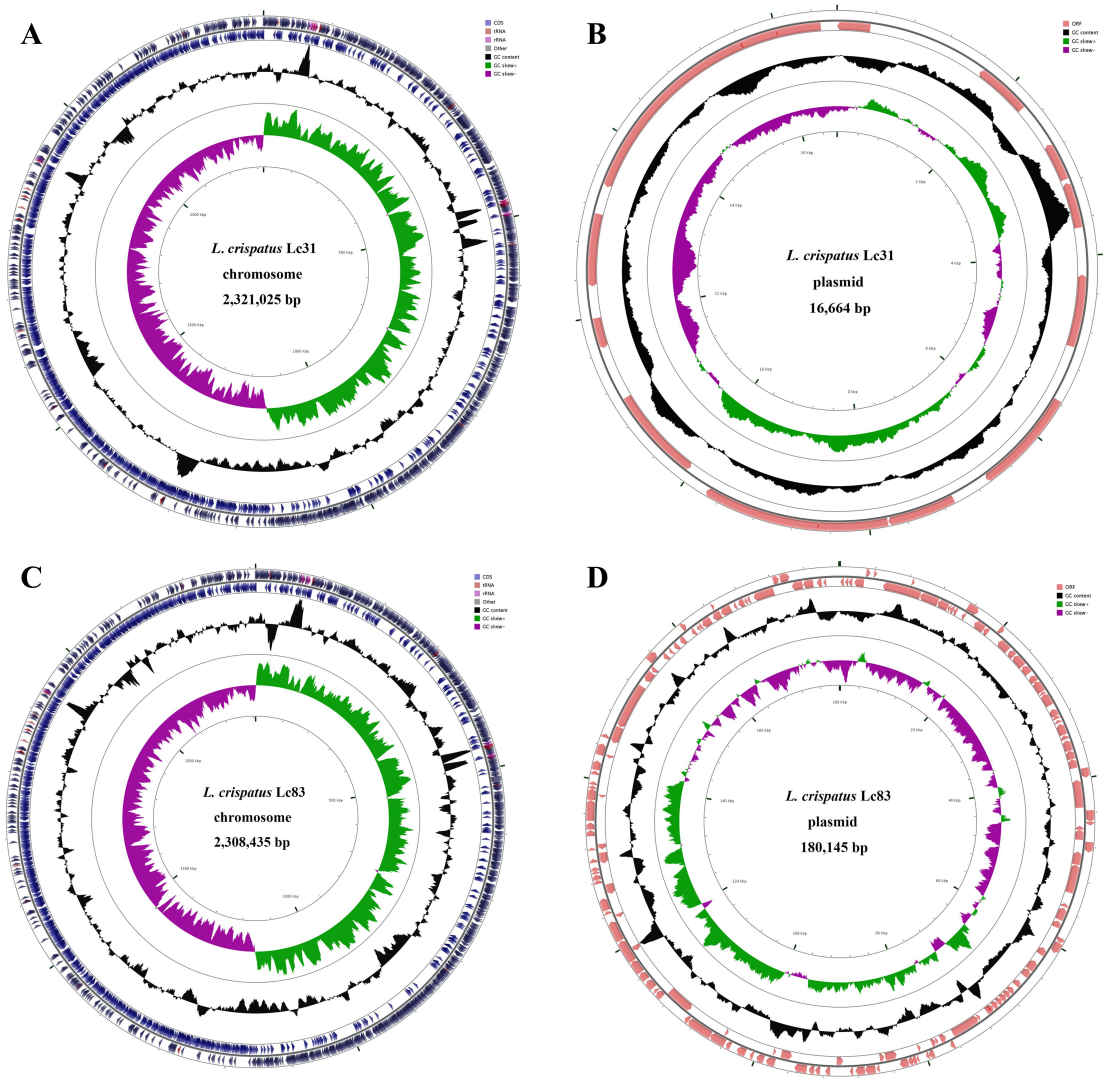

Figure 1 Circular map of the chromosome. (A) and plasmid (B) of *L. crispatus* Lc31 and the chromosome (C) and plasmid (D) of *L. crispatus* Lc83. For chromosomes, rings represent the following features from outer to inner: (1) genome size in Mb, (2) forward-strand genes colored according to cluster of orthologous groups (COG) classification, (3) reverse-strand genes colored according to COG classification, (4) forward-strand ncRNA, (5) reverse-strand ncRNA, (6) repeat, (7) GC content, and (8) GC skew. For plasmids, rings represent the following features from outer to inner: (1) genome size in kb, (2) forward-strand genes colored according to COG classification, (3) reverse-strand genes colored according to COG classification, (4) GC content, and (5) GC skew.

crispatus strains Lc31, AB70, FDAARGOS-743, PRL2021, 2029, PMC201, Lc116, Lc1226, Lc1700, and 1D contained type II-A CRISPR systems with a type II-A *cas* cluster adjacent to the CRISPR arrays. In the genomes of *L. crispatus* CO3MRSI1 and DC21.1, there were no *cas* genes next to CRISPR arrays.

Numerous GI and IS elements from various families were identified in the genomes of all 17 *L. crispatus* genital strains (Tables S7 and S8). The number of GIs in these *L. crispatus* strains ranged from 10 to 29 genital strains (Table S7). On average, more GIs were detected in the strains from the urogenital group (21.55 ± 3.47) than in those from the gastrointestinal group (17.83 ± 4.88), but the difference between the two groups was not statistically significant (Figure 3A). The distribution and abundance of IS elements in *L. crispatus* genomes were highly variable. Apart from the few IS elements located on the plasmids of *L. crispatus* Lc83, Lc116, and Lc1700, most IS elements were distributed on chromosomes rather than plasmids (Table S8). ISLh1 was the most abundant and widespread IS family in the genomes of the *L. crispatus* strains (Table S8). Significantly more IS elements were identified in the 11 strains of *L. crispatus* isolated from the urogenital tract (20.36 ± 8.46) than in the six strains isolated from the gastrointestinal tract (10.5 ± 9.59; Figure 3B). While the exact role of IS elements in the evolution of *L. crispatus* genomes remains poorly understood, such MGE elements can work as vehicles for horizontal gene transfer (HGT) due to their enhanced horizontal mobility. These numerous MGEs could allow genetic exchange between *L. crispatus* and other organisms and may be a major driver of HGT [54].

Consistent with the IS expansion in the genomes of urogenital strains of *L. crispatus*, significant differences in chromosome sizes and numbers of predicted CDSs were observed between strains isolated from different ecological niches (Table 1 and Figure 3C, D). Specifically, strains isolated from the gastrointestinal tract had a significantly smaller chromosome size and fewer CDSs than those isolated from the urogenital tract ($P < 0.05$). Additionally, urogenital

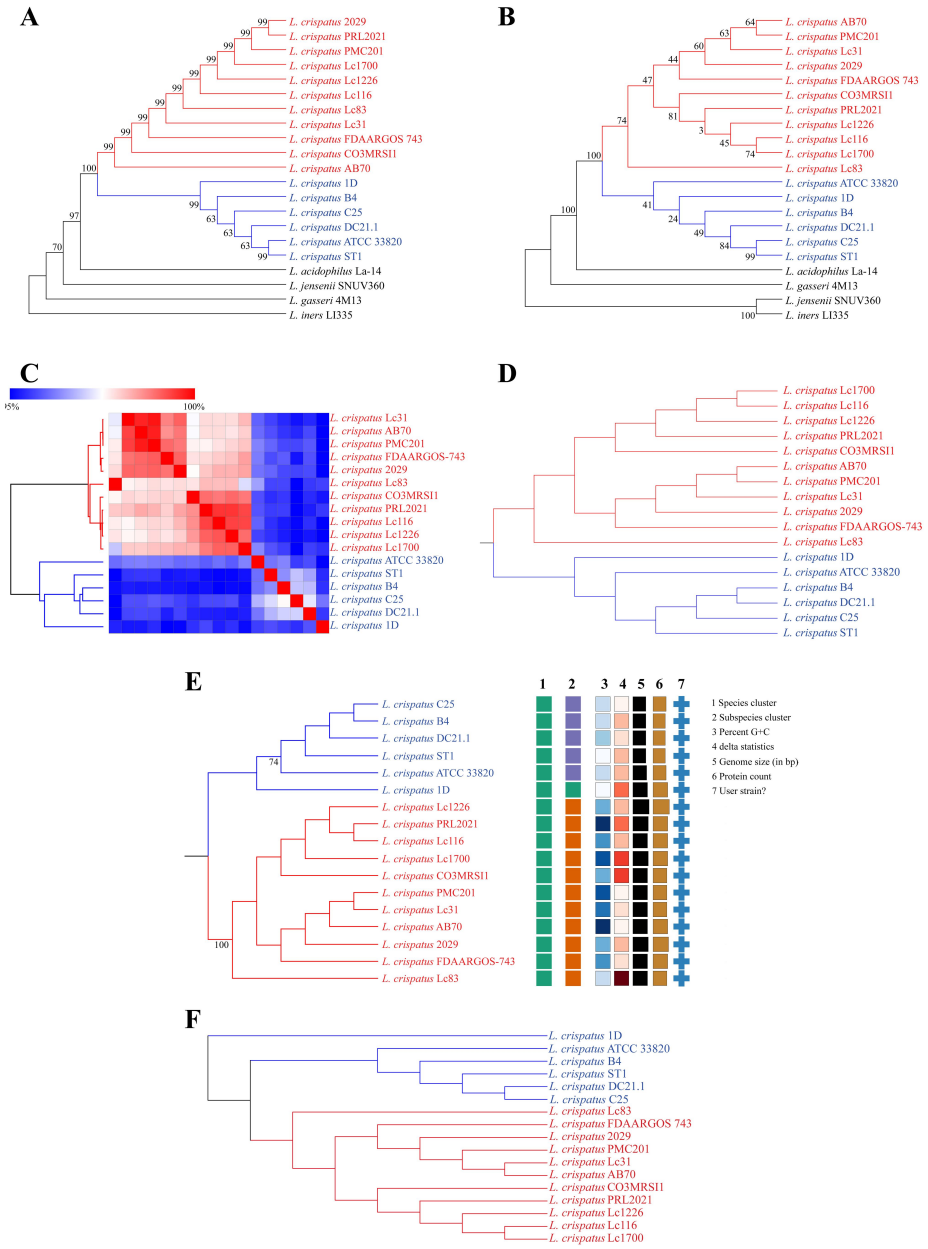

Figure 2 Phylogenetic trees based on *recA* gene sequences. (A) 31 conserved protein sequences (B) ANI matrix (C) CV approach (D) GBDP analysis (E) and core genes (F). The urogenital strains are marked in red, and the gastrointestinal strains are in blue.

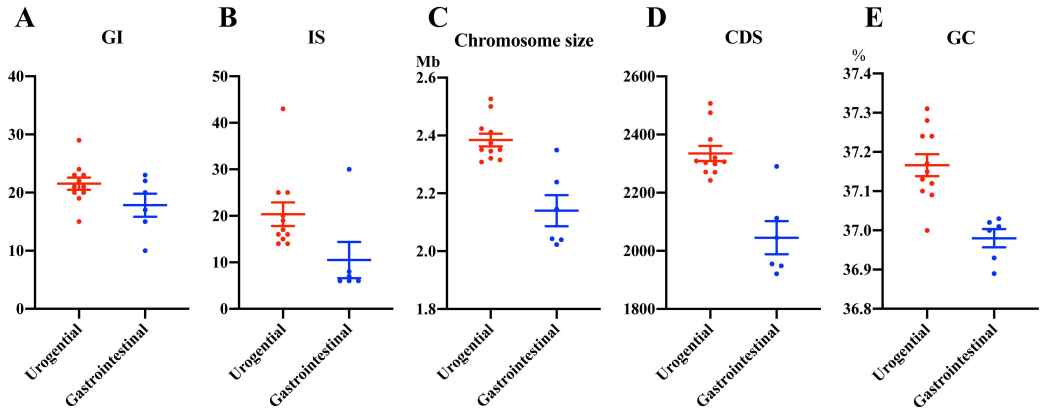

Figure 3 Comparison of (A) GIs, (B) ISs, (C) chromosome sizes, (D) CDS numbers and (E) GC content between urogenital and gastrointestinal strains of *L. crispatus* strains.

strains exhibited a higher genomic GC content than gastrointestinal strains (Table 1 and Figure 3E). It is worth noting that nucleotide composition in prokaryotic genomes can be shaped both by phylogeny and environment [55]. For certain phyla, their GC content can be determined by their living environments [55]. However, the exact environmental factor driving such variation in GC content in niche-specific strains of *L. crispatus* has yet to be determined.

BGCs, virulence and antibiotic resistance genes

Natural secondary metabolites produced by bacteria are important sources of antimicrobials and other bioactive compounds. BGCs were detected in 14 *L. crispatus* strains; the exceptions were *L. crispatus* B4, ST1, and DC21.1 (Table S9). Gasseriocin T (GT) was the most abundant bacteriocin type in these 14 strains, and the corresponding gene cluster was found in all 11 urogenital strains and in three of the six gastrointestinal strains. Additionally, the gene cluster for the biosynthesis of listeriolysin S (LLS) was present in seven urogenital strains and one gastrointestinal strain. GT was first isolated from *L. gasseri* and characterized as a heat-stable, pH-tolerant bacteriocin with wide antibacterial spectra against gram-positive foodborne pathogens [56]. LSS is a thiazole/oxazole-modified microcin produced by *Listeria monocytogenes* that promotes gut colonization of *L. monocytogenes* by killing specific gram-positive bacteria in a contact-dependent manner [57]. The antimicrobial spectrum and mechanism of these two bacteriocins have not been experimentally validated in *L. crispatus* thus far. However, it has been confirmed that lactic acid bacteria (LAB)-secreted bacteriocins have protective effects on GIT by eliminating pathogens and/or supporting the gut from bacterial colonization [58]. In addition, LAB-produced bacteriocins also promote urogenital tract health by their antimicrobial activity and inhibitory effects on biofilm formation [59]. In this study, all urogenital strains encoded at least one of the two bacteriocins, suggesting that the production of GT and LLS could favor *L. crispatus* colonization and exclusion of pathogens in urogenital tracts.

When considering the use of *L. crispatus* as a probiotic, the pathogenicity of individual strains should not be neglected. Therefore, the genomes of the 17 strains of *L. crispatus* examined in this study were screened for the presence of putative virulence genes. No VF genes were identified in any of these genomes. The occurrence and distribution of antibiotic resistance genes were subsequently investigated. A total of 12 antibiotic resistance genes were identified in the genomes of three GIT strains of *L. crispatus* (DC21.1, C25, and B4) and belonged to five categories: lincosamide, tetracycline, streptogramin, phenicol, and multidrug (Table S10).

Comparative genomic analysis

To assess the genetic diversity of 17 *L. crispatus* strains, the core genomes (1188 genes) and pangenomes (4006 genes) were identified using Figure 4A, B and Table S11 as references. The percentage of core genes in each genome ranged from 40.76% to 57.72%, indicating that approximately half of the genes in each genome were flexible. Except for PRL2021, unique genes were found in every genome. Among the other 16 strains, the number of unique genes varied from 3 (strain PMC201) to 249 (strain C25), with percentages ranging from 0.13% (strain PMC201) to 11.23% (strain C25) (Figure 4C). The unique genes were located on both chromosomes and plasmids. Among the 249 unique genes of *L. crispatus* C25, 58.23% were distributed on the plasmid, and 41.77% were chromosomal.

A COG functional analysis of the genes in the core and pangenomes was performed using Table 2 as a reference. The core genes were predominantly distributed in seven COG categories, namely, “translation, ribosomal structure, and biogenesis” [J], “replication, recombination, and repair” [L], “transcription” [K], “nucleotide transport and metabolism” [F], “carbohydrate transport and metabolism” [G], “inorganic ion transport and metabolism” [P], “cell wall/membrane/envelope biogenesis” [M], and “amino acid transport and metabolism” [E]. Collectively, these genes accounted for 52.19% of the core genome. The extent of pangenome expansion varied considerably across COG categories due to the uneven distributions of dispensable genes. The smallest expansions were observed in the COG categories “translation, ribosomal structure, and biogenesis” [J], “nucleotide transport and metabolism” [F], and “coenzyme transport and metabolism” [H], indicating that the functions and mechanisms of genes in these COG categories are more conserved. In contrast, the greatest expansion occurred in the COG categories “Cell motility” [N], “Secondary metabolite biosynthesis, transport, and catabolism” [Q], and “Defense mechanisms” [V], suggesting that the strains of *L. crispatus* have a variety of defense mechanisms to adapt to environmental stress.

Dispensable genes were also analyzed to determine whether their distribution was associated with the ecological niche of the strain. A total of 30 niche-specific genes were identified, including 13 that were exclusive to the six gastrointestinal strains and 17 shared only by the 11 urogenital strains (Table S12). All these genes were located on chromosomes, and none of them were detected on plasmids. COG annotations were also assigned to niche-specific genes, and the genes with assigned COG categories are listed in Table 3. Among the 13 genes shared by the gastrointestinal strains of *L. crispatus*, eight encoded poorly characterized proteins of unknown function. Three of the five genes with known functions encoded toxin–antitoxin proteins and were assigned to the COG category “Defense mechanisms” [V]. Seven of the 17 genes specific to urogenital strains of *L. crispatus* had known functions and were components of the *ula* cluster consisting of two units: an *ulaGCABDEF* operon and a divergent *ulaR* gene (Table 3 and Figure S6).

Many bacteria, both probiotics and pathogens, found in the human body can anaerobically catabolize L-ascorbate through factors encoded by the *ula* operon [54]. The *ulaCAB* genes are responsible for the uptake and phosphorylation of L-ascorbate through a phosphotransferase system. L-ascorbate 6-phosphate is generated intracellularly and is then transformed into 3-keto-L-gulonate 6-phosphate by L-ascorbate 6-phosphate lactonase (UlaG). The resulting 3-keto-L-gulonate 6-phosphate is converted to D-xylulose 5-phosphate through the sequential action of three catabolic enzymes, including UlaD (3-keto-L-gulonate 6-phosphate decarboxylase), UlaE (L-xylulose 5-phosphate 3-epimerase), and UlaF (L-ribulose 5-phosphate 4-epimerase). D-xylulose 5-phosphate is an important intermediate in the pentose phosphate pathway and can support bacterial growth when primary carbon sources other than L-ascorbate are limited [60].

L-Ascorbate (vitamin C) is an essential micronutrient in the human body with an antioxidant function in plasma. It has been found in high millimolar concentrations in cells and tissues but only low micromolar concentrations in plasma [61]. The presence of the *ula* operon in the urogenital strains of *L. crispatus* examined in this study suggests that L-ascorbate could serve as an alternative carbon source for bacteria in the urogenital tract of humans. Further research is necessary to determine the biological significance of L-ascorbate and its catabolism in these strains.

KEGG annotation was used to predict metabolic functions in the genomes. The KEGG map for pyruvate metabolism identified both L-lactate dehydrogenase and D-lactate dehydrogenase in the core genome, indicating that L- and D-lactate generation was present in all strains of *L. crispatus* (Figure S7). The L-ascorbate metabolism pathway was identified in the KEGG map for ascorbate and aldarate metabolism of the pangenome and the genomes of the 11 urogenital strains of *L. crispatus*, which was consistent with the COG analysis (Figure S8). KEGG annotation also identified a glycogen-degrading pullulanase (encoded by *pulA*) in the pangenome but not in the core genome (Figure S9 and S10). The presence of *pulA* was subsequently examined in the genomes of each of the 17 strains of *L. crispatus* by KEGG metabolic analysis. All strains, except AB70, ST1, and DC21.1, carried a copy of this gene.

CAZyme profiling

Mining of the 17 genomes of *L. crispatus* strains revealed the presence of 977 CAZyme-encoding genes (Table S13). In total, 537 glycoside hydrolases (GHs), 338 glycosyl transferases (GTs), 48 noncatalytic

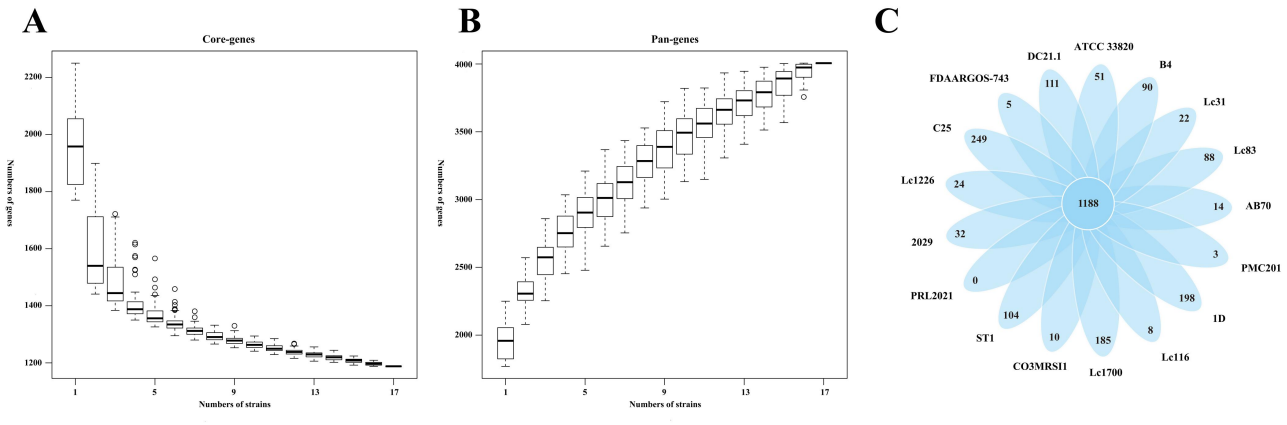

Figure 4 Core- and pangenome analysis encompassing the genomes of 17 strains of *L. crispatus*. (A) Plot showing the trajectory pattern of the reduction of the core-genome genes. (B) Plot showing the trajectory pattern of the expansion of pangenome genes. (C) Venn diagram showing the distribution of core and specific genes in the 17 strains of *L. crispatus*.

Table 2 Functional categories of pan, core-, dispensable-, and specific-genes in the genomes of the 17 strains of *L. crispatus* included in this study

COG Functional category	No. of pangenes	No. of core-genes	No. of dispensable-genes	No. of specific-genes
Total amount of input sequences	4006	1188	1624	1194
Total amount of scanned sequences	2878	1183	1150	545
Total number of sequences with assigned COG category	2484	1091	954	439
Total number of sequences with more than one COG category	90	37	36	17
Cellular Processes and Signaling				
[D] Cell cycle control, cell division, chromosome partitioning	43	25	10	8
[M] Cell wall/membrane/envelope biogenesis	168	56	61	51
[N] Cell motility	5	0	4	1
[O] Post-translational modification, protein turnover, and chaperones	39	27	9	3
[T] Signal transduction mechanisms	36	15	20	1
[U] Intracellular trafficking, secretion, and vesicular transport	42	24	11	7
[V] Defense mechanisms	93	20	51	22
Information Storage and Processing				
[J] Translation, ribosomal structure and biogenesis	159	148	10	1
[K] Transcription	250	79	112	59
[L] Replication, recombination and repair	258	86	116	56
Metabolism				
[C] Energy production and conversion	77	37	30	10
[E] Amino acid transport and metabolism	101	56	36	9
[F] Nucleotide transport and metabolism	96	71	22	3
[G] Carbohydrate transport and metabolism	194	65	59	70
[H] Coenzyme transport and metabolism	54	38	10	6
[I] Lipid transport and metabolism	52	21	25	6
[P] Inorganic ion transport and metabolism	100	59	32	9
[Q] Secondary metabolites biosynthesis, transport, and catabolism	12	2	6	4
Poorly Characterized				
[S] Function unknown	615	225	294	96

Table 3 Niche-specific genes with assigned COG categories

Gene ID[a]	Annotation	COG ID	Class ID	Class Function
Gastrointestinal tract-specific genes				
[ST1]_LCRIS_01765	Antitoxin component of the RelBE or YafQ-DinJ toxin-antitoxin module	COG3077	V	Defense mechanisms
[ST1]_LCRIS_01025	Protein N-acetyltransferase, RimJ/RimL family	COG1670	JO	Translation, ribosomal structure and biogenesis; Posttranslational modification, protein turnover, chaperones
[ST1]_LCRIS_01739	Antitoxin component of the RelBE or YafQ-DinJ toxin-antitoxin module	COG3077	V	Defense mechanisms
[ST1]_LCRIS_01996	DNA-binding transcriptional regulator, MerR family	COG0789	K	Transcription
[ST1]_LCRIS_01028	NTP pyrophosphatase, house-cleaning of noncanonical NTPs	COG1694	V	Defense mechanisms
Urogenital tract-specific genes				
[Lc31]_Lc31GL000343	Ascorbate-specific PTS system EIIC-type component UlaA	COG3037	G	Carbohydrate transport and metabolism
[Lc31]_Lc31GL000344	Phosphotransferase system, galactitol-specific IIB component	COG3414	G	Carbohydrate transport and metabolism
[Lc31]_Lc31GL000342	Phosphotransferase system mannitol/fructose-specific IIA domain (Ntr-type)	COG1762	GT	Carbohydrate transport and metabolism; Signal transduction mechanisms
[Lc31]_Lc31GL000345	3-keto-L-gulonate-6-phosphate decarboxylase	COG0269	G	Carbohydrate transport and metabolism
[Lc31]_Lc31GL000346	L-ribulose-5-phosphate 3-epimerase UlaE	COG3623	G	Carbohydrate transport and metabolism
[Lc31]_Lc31GL000348	Phosphoglycolate phosphatase, HAD superfamily	COG0546	C	Energy production and conversion
[Lc31]_Lc31GL000341	L-ascorbate metabolism protein UlaG, beta-lactamase superfamily	COG2220	G	Carbohydrate transport and metabolism

Table 4 McFarland's turbidity reached after 48 hrs incubation of *L. crispatus* Lc31 and Lc83 in minimal media supplemented with or without carbohydrates

	minimal medium (no substrate)	minimal medium (L-ascorbate)	minimal medium (glucose)	minimal medium (glycogen)
None[a]	0.02 ± 0.02	0.01 ± 0.01	0.02 ± 0.01	0.48 ± 0.02
Lc31	0.02 ± 0.02	0.24 ± 0.01	> 4.00	3.47 ± 0.02
Lc83	0.01 ± 0.01	0.20 ± 0.02	> 4.00	1.16 ± 0.02

carbohydrate-binding modules (CBMs), 38 carbohydrate esterases (CEs), and 16 proteins with auxiliary activity (AA) were identified. Further analysis revealed significantly different compositions of CAZymes between urogenital and gastrointestinal strains of *L. crispatus*. Urogenital strains had significantly more CEs ($P = 0.044$) but significantly fewer GTs ($P = 0.048$) than gastrointestinal strains (Table S13). Specifically, the number of copies of CAZyme-encoding genes also varied between the two groups. Urogenital strains had significantly more copies of genes in the CE9 family ($P = 0.035$), while gastrointestinal strains had significantly more copies of genes in the GH1, GH8, GH13, GH36, GH68, and GT2 families ($P = 0.012$, $P = 0.035$, $P = 0.003$, $P = 0.002$, $P = 0.009$, and $P = 0.008$, respectively). In addition, the CAZyme gene families GH20, GH38, GH92, and GH125 were unique to the urogenital strains, while the GH3, GH33, GH140 and GT113 families were detected only in gastrointestinal strains but not urogenital strains. In addition, there were differences in the predicted substrate specificity of the abovementioned CAZymes. The predicted substrates of the CE9, GH20, and GH125 families abundant in urogenital strains were O-linked and N-linked host glycans, while the predicted substrates of the GH1, GH8, GH13, and GH68 families from gastrointestinal strains were lactose, β-glucan, starch, and fructan, respectively [62]. Collectively, these results indicated that there were significant differences in carbohydrate metabolism between the urogenital and gastrointestinal groups of *L. crispatus*. This difference could be related to their niches and the environmental carbohydrate composition and availability that drive niche adaptation.

Growth on carbohydrate substrates

To investigate whether L-ascorbate could support the growth of *L. crispatus* strains Lc31 and Lc83, we examined the capacity of these strains for *in vitro* fermentation of different carbohydrate substrates. The negative control (without carbohydrate) showed no growth, while the turbidity value markedly increased in the presence of L-ascorbate (Table 4), indicating that this substrate supported *L. crispatus* growth. These findings are consistent with the results of the comparative genomic analysis described above. Both genomes of urogenital *L. crispatus* strains, Lc31 and Lc83, carry *ula* operons that encode the

necessary enzymes in the L-ascorbate catabolizing pathway. In addition, a *pulA* gene, which encodes the glycogen-degrading pullulanase, was also identified in the genomes of *L. crispatus* Lc31 and Lc83. These two strains exhibited growth on glycogen, but optimal growth was observed when glucose was used as the sole carbon source.

Antibiotic susceptibility

Although no antibiotic resistance gene was identified in the genomes of *L. crispatus* Lc31 and Lc83 by genome data mining, we also performed experiments to determine the susceptibility of these two *L. crispatus* strains to 18 antibiotics, and the results are shown in Table S14. Both strains of *L. crispatus* exhibited sensitivity to tetracyclines, macrolides, cephalosporins, glycopeptides, oxazolidinones, carbapenems, penicillins, clindamycin and chloramphenicol but not to aminoglycosides and quinolones, which was consistent with the behavior of other lactic acid bacteria [63].

Discussion

L. crispatus is a potential probiotic with anti-cervical cancer activity, but the underlying mechanisms remain unclear. The strains of *L. crispatus* occupy multiple ecological niches, including the UGT and GIT of humans, as well as the GIT of animals and poultry [10, 11]. It is uncertain whether the nature of *L. crispatus* strains from different ecological niche sources is different, which is of great significance for elucidating the mechanism and guiding the development of functional strains.

In this study, we isolated two *L. crispatus* strains (Lc31 and Lc83) from the vaginas of healthy women of reproductive age, obtained their complete genome sequences and compared them with 15 other strains of *L. crispatus* with complete genome sequences. Although hundreds of *L. crispatus* genome sequences have been deposited in the NCBI database, most of them are incomplete genomes at the contig or scaffold level. The value of phylogenetic and functional genomic results was limited by the incompleteness of the genome sequences. Therefore, only 15 complete genomes were selected and analyzed in this study.

A set of phylogenetic and phylogenomic methods were used to explore the intraspecies diversity among the strains of *L. crispatus*. The results of 16S rRNA gene sequence, dDDH and ANI analyses supported the classification of all these strains as *L. crispatus*. In addition, phylogenetic trees based on the *recA* gene and MLSA showed that urogenital and gastrointestinal strains of *L. crispatus* clustered separately. Four phylogenomic approaches based on whole-genome sequences, which can provide sufficient resolution at the species level and below, were used to verify the results obtained from the phylogenetic analysis. Both phylogenetic and phylogenomic results in this study supported the separate clustering of urogenital and gastrointestinal strains of *L. crispatus,* as reported in two previous studies [8, 11]. This niche-dependent clustering suggests that there may be differences in traits between the two clusters of *L. crispatus* strains.

Comparative genomic analysis tools were then applied to illustrate the underlying genetic mechanisms of niche-specific adaptation in *L. crispatus* strains and explore the potential anticancer characteristics of UGT strains. First, numerous MGEs were detected and compared in the genomes of *L. crispatus* strains. On average, more GIs and ISs were detected in the strains from the UGT group than in those from the GIT group. These numerous MGEs can allow genetic exchange between *L. crispatus* and other organisms sharing the same living environments and promote niche-specific evolution. Consistently, higher GC content, larger chromosome sizes and more CDSs were observed in the strains from the UGT group. This could indicate more active HGT events between *L. crispatus* and other organisms in the urogenital tracts of humans.

To explore the ecological significance of HGT in facilitating adaptation of *L. crispatus* strains to different ecological niches, we conducted a search for the composition of 23 BGCs in Table S9. We found that, except for the five GT gene clusters in the genomes of *L. crispatus* strains Lc83, PRL2021, Lc116, C25, and 1D, all 18 BGC gene clusters in proximity to bacteriocin biosynthetic-related genes in the other strains encoded transposases. Through adjacent gene analysis of the 12 antibiotic resistance genes encoded by the three strains in Table S10, we discovered that, except for two tetracycline resistance genes, all other 10 resistance genes for macrolides, lincosamides, streptogramins, and phenicols had transposase-encoding genes in their respective genomic regions. These findings suggest that both the resistance genes and BGCs in *L. crispatus* genomes are acquired through HGT. Based on this evidence, we can infer that *L. crispatus* strains residing in the urogenital tract are more likely to obtain BGC genes via HGT from bacteria of the same or other species cohabiting in the urogenital tract. On the other hand, *L. crispatus* strains inhabiting the gastrointestinal tract tend to acquire antibiotic resistance genes from other intestinal bacteria via HGT, allowing them to thrive in the antibiotic-rich digestive environment. Therefore, HGT may be an important mechanism for *L. crispatus* strains in different ecological niches to gain a survival advantage in their respective environments.

Second, a total of 30 niche-specific genes were identified by comparative genomic analysis, including 13 that were exclusive to the GIT strains and 17 shared only by the 11 UGT strains. Seven of the 17 genes specific to urogenital strains of *L. crispatus* had known functions and were components of the *ula* cluster consisting of two units: an *ulaGCABDEF* operon and a divergent *ulaR* gene. The *ula* operon encodes a set of enzymes that are responsible for anaerobically catabolizing L-ascorbate [60]. The presence of the *ula* operon in the urogenital strains of *L. crispatus* suggests that L-ascorbate could serve as an alternative carbon source for bacteria in the urogenital tract of humans. Furthermore, *in vitro* fermentation was performed and showed that L-ascorbate supported the growth of *L. crispatus* strains Lc31 and Lc83, which is consistent with the results of the comparative genomic analysis described above.

Third, when considering the use of *L. crispatus* as an anticancer probiotic, the safety of individual strains should not be neglected. Genome data mining showed that all UGT strains encoded at least one bacteriocin, which could favor *L. crispatus* colonization and exclusion of pathogens in the urogenital tract. In addition, no VF genes were identified in any of these genomes. The universal presence of genes encoding bacteriocins and the absence of virulence factors in urogenital strains support the possibility of *L. crispatus* as a potential urogenital probiotic. In addition, no antibiotic resistance gene was detected in the genomes of the UGT strains of *L. crispatus*. Antibiotic susceptibility tests were also carried out in *L. crispatus* Lc31 and Lc83 and showed that these two UGT strains exhibited sensitivity to most of the common antibiotics except aminoglycosides and quinolones, which was consistent with the behavior of other lactic acid bacteria and can be used as a safe probiotic [63].

In conclusion, the complete genomes of two strains of *L. crispatus* (Lc31 and Lc83) isolated from the vaginas of healthy women of reproductive age were sequenced and characterized. Phylogenetic and phylogenomic analyses of the two isolates and 15 other strains of *L. crispatus* with complete genomes revealed that strains from the two major niches, UGT and GIT, clustered separately in the phylogenetic trees. Strains from the UGT had a larger genome size, higher GC content, and more CDSs and IS elements compared with those from the GIT, indicating the likelihood of active HGT occurring in this niche. The universal presence of genes encoding bacteriocins and the absence of virulence factors in urogenital strains support the possibility of *L. crispatus* as a potential anticancer probiotic. Comparative genomic analysis identified an *ula* gene cluster exclusively responsible for L-ascorbate utilization in strains isolated from urogenital tracts. Carbohydrate fermentation experiments confirmed that L-ascorbate supported the growth of *L. crispatus* Lc31 and Lc83, consistent with the results of comparative genomic analysis. Additionally, a *pulA* gene encoding the glycogen-degrading pullulanase was identified in the genomes of *L. crispatus* Lc31 and Lc83. These two strains exhibited growth on glycogen, but optimal growth was observed when glucose was used as the sole carbon

source. In addition, there were no strain-specific phenotypes or genes of antibiotic resistance in *L. crispatus* Lc31 and Lc83.

In future research on probiotics targeting women's urogenital health, it is advisable to utilize *L. crispatus* strains from the same ecological niche to fully leverage their inherent adaptability to specific environments, such as their ability to utilize L-ascorbate for growth. Another advantage of using strains from the same source is that they carry and encode beneficial metabolites, such as bacteriocins, that contribute to their own survival and improve the microenvironment. Hence, our findings enhance the comprehension of how the genome influences niche adaptation in *L. crispatus* and establish a basis for investigating the mechanisms by which urogenital-derived *L. crispatus* promotes female health.

References

1. Łaniewski P, Ilhan ZE, Herbst-Kralovetz MM. The microbiome and gynaecological cancer development, prevention and therapy. *Nat Rev Urol* 2020;17(4):232–250. Available at: http://doi.org/10.1038/s41585-020-0286-z
2. Ravel J, Gajer P, Abdo Z, et al. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci USA* 2010;108(Suppl 1):4680–4687. Available at: http://doi.org/10.1073/pnas.1002611107
3. Wang H, Ma Y, Li R, Chen X, Wan L, Zhao W. Associations of Cervicovaginal Lactobacilli With High-Risk Human Papillomavirus Infection, Cervical Intraepithelial Neoplasia, and Cancer: A Systematic Review and Meta-Analysis. *J Infect Dis* 2019;220(8):1243–1254. Available at: http://doi.org/10.1093/infdis/jiz325
4. Petrova MI, Reid G, Vaneechoutte M, Lebeer S. Lactobacillus iners : Friend or Foe? *Trends Microbiol* 2017;25(3):182–191. Available at: http://doi.org/10.1016/j.tim.2016.11.007
5. Liu Y, Xu C, Pan J, Sun C, Zhou H, Meng Y. Significance of the viral load of high-risk HPV in the diagnosis and prediction of cervical lesions: a retrospective study. *BMC Womens Health* 2021;21(1):353. Available at: http://doi.org/10.1186/s12905-021-01493-0
6. So KA, Yang EJ, Kim NR, et al. Changes of vaginal microbiota during cervical carcinogenesis in women with human papillomavirus infection. *PLoS ONE* 2020;15(9):e0238705. Available at: http://doi.org/10.1371/journal.pone.0238705
7. Norenhag J, Du J, Olovsson M, Verstraelen H, Engstrand L, Brusselaers N. The vaginal microbiota, human papillomavirus and cervical dysplasia: a systematic review and network meta-analysis. *BJOG* 2020;127(2):171–180. Available at: http://doi.org/10.1111/1471-0528.15854
8. Mancabelli L, Mancino W, Lugli GA, et al. Comparative Genome Analyses of Lactobacillus crispatus Isolates from Different Ecological Niches Reveal an Adaptation of This Species to the Human Vaginal Environment. *Appl Environ Microbiol* 2021;87(8):e02899–20. Available at: http://doi.org/10.1128/AEM.02899-20
9. Fontana F, Alessandri G, Lugli GA, et al. Probiogenomics Analysis of 97 Lactobacillus crispatus Strains as a Tool for the Identification of Promising Next-Generation Probiotics. *Microorganisms* 2020;9(1):73. Available at: http://doi.org/10.3390/microorganisms9010073
10. Pan M, Hidalgo-Cantabrana C, Barrangou R. Host and body site-specific adaptation of Lactobacillus crispatus genomes. *NAR Genom Bioinform* 2020;2(1):lqaa001. Available at: http://doi.org/10.1093/nargab/lqaa001
11. Zhang Q, Zhang L, Ross P, Zhao J, Zhang H, Chen W. Comparative Genomics of Lactobacillus crispatus from the Gut and Vagina Reveals Genetic Diversity and Lifestyle Adaptation. *Genes (Basel)* 2020;11(4):360. Available at: http://doi.org/10.3390/genes11040360
12. Chen C, Yu L, Tian F, Zhao J, Zhai Q. Identification of Novel Bile Salt-Tolerant Genes in Lactobacillus Using Comparative Genomics and Its Application in the Rapid Screening of Tolerant Strains. *Microorganisms* 2022;10(12):2371. Available at: http://doi.org/10.3390/microorganisms10122371
13. Wang SY, Chen YP, Huang RF, et al. Subspecies Classification and Comparative Genomic Analysis of Lactobacillus kefiranofaciens HL1 and M1 for Potential Niche-Specific Genes and Pathways. *Microorganisms* 2022;10(8):1637. Available at: http://doi.org/10.3390/microorganisms10081637
14. Son S, Lee R, Park SM, et al. Complete genome sequencing and comparative genomic analysis of Lactobacillus acidophilus C5 as a potential canine probiotics. *J Anim Sci Technol* 2021;63(6):1411–1422. Available at: http://doi.org/10.5187/jast.2021.e126
15. Myers EW, Sutton GG, Delcher AL, et al. A Whole-Genome Assembly ofDrosophila. *Science* 2000;287(5461):2196–2204. Available at: http://doi.org/10.1126/science.287.5461.2196
16. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20(9):1297–1303. Available at: http://doi.org/10.1101/gr.107524.110
17. Li R, Yu C, Li Y, et al. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 2009;25(15):1966–1967. Available at: http://doi.org/10.1093/bioinformatics/btp336
18. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 2007;23(6):673–679. Available at: http://doi.org/10.1093/bioinformatics/btm009
19. Chan PP, Lowe TM. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods Mol Biol* 2019:1–14. Available at: http://doi.org/10.1007/978-1-4939-9173-0_1
20. Lagesen K, Hallin P, Rødland EA, Stærfeldt H-H, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 2007;35(9):3100–3108. Available at: http://doi.org/10.1093/nar/gkm160
21. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016;44(D1):D457–D462. Available at: http://doi.org/10.1093/nar/gkv1070
22. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res* 2015;43(D1):D261–D269. Available at: http://doi.org/10.1093/nar/gku1223
23. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium *Nat Genet* 2000;25(1):25–29. Available at: http://doi.org/10.1038/75556
24. Huerta-Cepas J, Forslund K, Coelho LP, et al. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol Biol Evol* 2017;34(8):2115–2122. Available at: http://doi.org/10.1093/molbev/msx148
25. Consortium U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;47 (D1):D506–D515. Available at: http://doi.org/10.1093/nar/gky1049
26. Lee I, Chalita M, Ha SM, Na SI, Yoon SH, Chun J. ContEst16S: an algorithm that identifies contaminated prokaryotic genomes using 16S RNA gene sequences. *Int J Syst Evol Microbiol* 2017;67(6):2053–2057. Available at: http://doi.org/10.1099/ijsem.0.001872
27. Ellegaard KM, Engel P. Beyond 16S rRNA Community Profiling: Intra-Species Diversity in the Gut Microbiota. *Front Microbiol* 2016;7:1475. Available at:

http://doi.org/10.3389/fmicb.2016.01475
28. Singh S, Goswami P, Singh R, Heller KJ. Application of molecular identification tools for Lactobacillus, with a focus on discrimination between closely related species: A review. *LWT-Food Sci Technol* 2009;42(2):448–457. Available at: http://doi.org/10.1016/j.lwt.2008.05.019
29. Felis GE, Dellaglio F, Mizzi L, Torriani S. Comparative sequence analysis of a recA gene fragment brings new evidence for a change in the taxonomy of the Lactobacillus casei group. *Int J Syst Evol Microbiol* 2001;51(Pt 6):2113–2117. Available at: http://doi.org/10.1099/00207713-51-6-2113
30. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;32(5):1792–1797. Available at: http://doi.org/10.1093/nar/gkh340
31. Stecher G, Tamura K, Kumar S. Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Mol Biol Evol* 2020;37(4):1237–1239. Available at: http://doi.org/10.1093/molbev/msz312
32. Kerepesi C, Bánky D, Grolmusz V. AmphoraNet: The webserver implementation of the AMPHORA2 metagenomic workflow suite. *Gene* 2014;533(2):538–540. Available at: http://doi.org/10.1016/j.gene.2013.10.015
33. Meier-Kolthoff JP, Carbasse JS, Peinado-Olarte RL, Göker M. TYGS and LPSN: a database tandem for fast and reliable genome-based classification and nomenclature of prokaryotes. *Nucleic Acids Res* 2021;50(D1):D801–D807. Available at: http://doi.org/10.1093/nar/gkab902
34. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci USA* 2009;106(45):19126–19131. Available at: http://doi.org/10.1073/pnas.0906412106
35. Zuo G, Hao B. CVTree3 Web Server for Whole-genome-based and Alignment-free Prokaryotic Phylogeny and Taxonomy. *Genomics Proteomics Bioinformatics* 2015;13(5):321–331. Available at: http://doi.org/10.1016/j.gpb.2015.08.004
36. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* 2021;49(W1):W293–W296. Available at: http://doi.org/10.1093/nar/gkab301
37. Darling AE, Mau B, Perna NT. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE* 2010;5(6):e11147. Available at: http://doi.org/10.1371/journal.pone.0011147
38. Tesler G. GRIMM: genome rearrangements web server. *Bioinformatics* 2002;18(3):492–493. Available at: http://doi.org/10.1093/bioinformatics/18.3.492
39. Noe L, Kucherov G. YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res* 2005;33(Web Server issue):W540–W543. Available at: http://doi.org/10.1093/nar/gki478
40. Arndt D, Grant JR, Marcu A, et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016;44(W1):W16–W21. Available at: http://doi.org/10.1093/nar/gkw387
41. Couvin D, Bernheim A, Toffano-Nioche C, et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res* 2018;46(W1):W246–W251. Available at: http://doi.org/10.1093/nar/gky425
42. Bertelli C, Laird MR, Williams KP, et al. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res* 2017;45(W1):W30–W35. Available at: http://doi.org/10.1093/nar/gkx343
43. Johansson MHK, Bortolaia V, Tansirichaiya S, Aarestrup FM, Roberts AP, Petersen TN. Detection of mobile genetic elements associated with antibiotic resistance in Salmonella enterica using a newly developed web tool: MobileElementFinder. *J Antimicrob Chemother* 2021;76(1):101–109. Available at: http://doi.org/10.1093/jac/dkaa390
44. Blin K, Shaw S, Kloosterman AM, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res* 2021;49(W1):W29–W35. Available at: http://doi.org/10.1093/nar/gkab335
45. Liu B, Zheng D, Jin Q, Chen L, Yang J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res* 2019;47(D1):D687–D692. Available at: http://doi.org/10.1093/nar/gky1080
46. Alcock BP, Raphenya AR, Lau TTY, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2019;48(D1):D517–D525. Available at: http://doi.org/10.1093/nar/gkz935
47. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010;26(5):680–682. Available at: http://doi.org/10.1093/bioinformatics/btq003
48. Zhang H, Yohe T, Huang L, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* 2018;46(W1):W95–W101. Available at: http://doi.org/10.1093/nar/gky418
49. Morishita T, Deguchi Y, Yajima M, Sakurai T, Yura T. Multiple nutritional requirements of lactobacilli: genetic lesions affecting amino acid biosynthetic pathways. *J Bacteriol* 1981;148(1):64–71. Available at: http://doi.org/10.1128/jb.148.1.64-71.1981
50. Sharma P, Tomar SK, Sangwan V, et al. Antibiotic Resistance of *Lactobacillus sp.* Isolated from Commercial Probiotic Preparations. *J Food Safety* 2016;36(1):38–51. Available at: https://doi.org/10.1111/jfs.12211
51. Tindall BJ, Rosselló-Móra R, Busse HJ, Ludwig W, Kämpfer P. Notes on the characterization of prokaryote strains for taxonomic purposes. *Int J Syst Evol Microbiol* 2010;60(Pt 1):249–266. Available at: http://doi.org/10.1099/ijs.0.016949-0
52. Yolken RH, Severance EG, Sabunciyan S, et al. Metagenomic Sequencing Indicates That the Oropharyngeal Phageome of Individuals With Schizophrenia Differs From That of Controls. *Schizophr Bull* 2015;41(5):1153–1161. Available at: http://doi.org/10.1093/schbul/sbu197
53. Koonin EV, Makarova KS, Zhang F. Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol* 2017;37:67–78. Available at: http://doi.org/10.1016/j.mib.2017.05.008
54. Eraclio G, Ricci G, Fortina MG. Insertion sequence elements in Lactococcus garvieae. *Gene* 2015;555(2):291–296. Available at: http://doi.org/10.1016/j.gene.2014.11.019
55. Reichenberger ER, Rosen G, Hershberg U, Hershberg R. Prokaryotic Nucleotide Composition Is Shaped by Both Phylogeny and the Environment. *Genome Biol Evol* 2015;7(5):1380–1389. Available at: http://doi.org/10.1093/gbe/evv063
56. Kawai Y, Saitoh B, Takahashi O, et al. Primary Amino Acid and DNA Sequences of Gassericin T, a Lactacin F-Family Bacteriocin Produced byLactobacillus gasseriSBT2055. *Biosci Biotechnol Biochem* 2000;64(10):2201–2208. Available at: http://doi.org/10.1271/bbb.64.2201
57. Meza-Torres J, Lelek M, Quereda JJ, et al. Listeriolysin S: A bacteriocin from Listeria monocytogenes that induces membrane permeabilization in a contact-dependent manner. *Proc Natl Acad Sci USA* 2021;118(40):e2108155118. Available at: http://doi.org/10.1073/pnas.2108155118
58. Alvarez-Sieiro P, Montalbán-López M, Mu D, Kuipers OP. Bacteriocins of lactic acid bacteria: extending the family. *Appl*

Microbiol Biotechnol 2016;100(7):2939–2951. Available at: http://doi.org/10.1007/s00253-016-7343-9

59. Khan F, Tabassum N, Kim YM. A strategy to control colonization of pathogens: embedding of lactic acid bacteria on the surface of urinary catheter. *Appl Microbiol Biotechnol* 2020;104(21):9053–9066. Available at: http://doi.org/10.1007/s00253-020-10903-6
60. Linares D, Michaud P, Delort AM, Traïkia M, Warrand J. Catabolism of L-ascorbate by Lactobacillus rhamnosus GG. *J Agric Food Chem* 2011;59(8):4140–4147. Available at: http://doi.org/10.1021/jf104343r
61. Injarabian L, Scherlinger M, Devin A, Ransac S, Lykkesfeldt J, Marteyn BS. Ascorbate maintains a low plasma oxygen level. *Sci Rep* 2020;10(1):10659. Available at: http://doi.org/10.1038/s41598-020-67778-w
62. Wang W, Hu H, Zijlstra RT, Zheng J, Gänzle MG. Metagenomic reconstructions of gut microbial metabolism in weanling pigs. *Microbiome* 2019;7(1):48. Available at: http://doi.org/10.1186/s40168-019-0662-1
63. Fan X, Jiang X, Guo Y, et al. In vitro and in vivo evaluation of the safety of Levilactobacillus brevis CGMCC1.5954 with probiotic potential based on tri-generation whole genome sequencing and animal studies. *Food Biosci* 2023;53:102654. Available at: http://doi.org/10.1016/j.fbio.2023.102654