Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Strain-resolved comparative genomics and functional characterization of culturable lactic acid bacteria from Guizhou sour soup

fan bai
Guiyang Medical University
nanlan wang
Guiyang Medical University
yang liu
Guiyang Medical University
qian wen
Guiyang Medical University
canying hu
Academy of Military Medical Sciences
siqi zhu
Harbin Medical University
rong mou
Guiyang Medical University
ruifu yang
Academy of Military Medical Sciences
yanpin han
hypiota@hotmail.com
Academy of Military Medical Sciences
peng luo
Guiyang Medical University

Research Article

Keywords: lactic acid bacteria, comparative genomics, pan-genome, fermented food, cholesterol removal, strain selection

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10810674/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Guizhou sour soup is a traditional fermented food and a reservoir of diverse lactic acid bacteria (LAB), yet the strain-resolved genomic architecture and functional heterogeneity of its culturable LAB remain poorly defined. We established a large culture collection and integrated comparative genomics with targeted analyses of bile salt hydrolase (BSH), exopolysaccharide/capsular polysaccharide (EPS/CPS) loci, genomic safety, and in vitro phenotypes to evaluate strain-level diversity and identify traits relevant to functional strain selection. Among 1,741 culturable isolates recovered from red and white sour soups, Lactobacillales accounted for 85.69% and 95.24%, respectively. Whole-genome sequencing of 172 isolates, comprising 92 *Lacticaseibacillus paracasei*, 30 *Lactiplantibacillus plantarum*, and 50 *Limosilactobacillus fermentum*, revealed open pan-genomes with distinct species-level functional architectures. *L. paracasei* showed the greatest accessory-genome contribution, *L. plantarum* the highest mean normalized BSH representation, and *L. fermentum* distinct oxidative-stress- and amino-acid-associated profiles. No antimicrobial resistance or virulence-associated genes were detected at the applied screening thresholds. BSH analyses showed lineage-specific sequence variation but conserved catalytic Cys2 and ligand-proximal architecture, whereas EPS/CPS loci varied among species and strains. Cholesterol removal was strongly strain- and condition-dependent: *L. paracasei* R1156 reached 71.69 ± 4.84% in standard MRS, while *L. plantarum* R2345, R2037, R1621, and R203 maintained 56.85–64.32% removal under 0.1% bile salts, exceeding the reference strain *Lacticaseibacillus rhamnosus* GG (LGG). These findings establish a strain-resolved genomic and phenotypic resource and demonstrate the value of integrating comparative genomics with direct phenotypic screening for rational selection of food-associated LAB for further biotechnological evaluation.

1. Introduction

Traditional fermented foods are dynamic microbial ecosystems and important reservoirs of food-associated microorganisms with diverse genomic and phenotypic traits (Marco et al. 2021; Mukherjee et al. 2024). Recent food-microbiome research has increasingly moved from species-level profiling toward strain-resolved characterization, because fermentation-relevant and host-associated traits may vary substantially within a species (Carlino et al. 2024; Tian et al. 2025). This shift is particularly relevant for regional fermented foods, where culture-based recovery not only captures living microbial diversity but also links genomic variation to experimentally testable phenotypes and preserves strains for downstream technological evaluation, as recently demonstrated in tomato sour soup from Guizhou (Zhou et al. 2025). Comparative genomic work in food-associated lactobacilli has likewise shown that open pan-genomes can coexist with lineage- and source-associated functional differentiation (Mejía-Caballero et al. 2025). Despite growing interest in widely studied fermented foods such as kimchi, dairy products, and fermented vegetables (Carlino et al. 2024), few studies have combined large-scale culture collections, strain-resolved comparative genomics, and targeted functional validation across multiple co-occurring LAB species from the same fermented-food ecosystem. Such an integrated framework is important for distinguishing broadly conserved traits from strain-specific features relevant to functional strain selection (Kahraman Ilıkkan 2025).

Bile-associated metabolism and cholesterol removal provide a tractable phenotypic axis for evaluating functional variation among food-associated LAB. Proposed mechanisms of LAB-associated cholesterol removal include BSH-mediated deconjugation, co-precipitation with deconjugated bile acids, cell-surface adsorption, and cellular assimilation (Tjandrawinata et al. 2022). Substantial strain-level differences in cholesterol assimilation and stress tolerance have also been reported among LAB recovered from Chinese fermented rice (Qayyum et al. 2023), consistent with the broader view that cholesterol-related phenotypes depend on both strain background and culture environment. Red and white Guizhou sour soups differ in raw materials and microbial composition (Zhou et al. 2026), providing an ecological context in which the contributions of genomic background and culture condition to strain-level functional variation can be examined.

In this study, we established a large-scale culturable LAB resource from Guizhou sour soup and integrated comparative pan-genomics with targeted phenotypic evaluation of phylogenetically and genomically representative strains. We aimed to characterize culturable microbial recovery from red and white sour soups, define species- and strain-level genomic and functional variation in three focal LAB species, assess preliminary genome-based safety, and determine the extent to which selected genomic features corresponded to gastrointestinal tolerance and cholesterol-removal performance under defined in vitro conditions. By combining genome-informed prioritization with direct phenotypic testing, we sought to provide a rational framework for selecting food-associated LAB for further functional and biotechnological evaluation.

2. Materials and Methods

2.1 Sample Collection, Strain Isolation, and Identification

Eight traditional sour soup samples, comprising four red sour soups (R1–R4) and four white sour soups (B1–B4), were obtained from local vendors in Kaili, Guizhou Province, China, using aseptic procedures. The samples were transported to the laboratory at 4°C and processed upon arrival. For anaerobic enrichment, 0.5 mL of each sour soup sample was inoculated into an anaerobic blood-culture bottle (Autobio, China) containing 5 mL of commercially obtained rumen fluid (FeiMoBio Life Sciences, Hunan, China) and 5 mL of goat blood (Solarbio, Beijing, China). The enrichment cultures were incubated anaerobically at 37°C for 15 d.

Microbial isolation was performed on days 0, 3, 6, 9, 12, and 15. At each time point, samples were serially diluted from 10^{-3} to 10^{-9} in sterile phosphate-buffered saline (PBS). Aliquots (100 µL) were spread onto Columbia blood agar and incubated anaerobically at 37°C for 48 h. Colonies with distinct morphologies were selected and repeatedly streaked until pure cultures were obtained. Purified isolates were propagated in de Man, Rogosa, and Sharpe (MRS) broth.

Preliminary identification was performed using a Zybio EXS2000 matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) system at Beijing Yujing

Pharmaceutical Co., Ltd. (Beijing, China). Fresh colonies were smeared directly onto a MALDI target plate, treated with 1 μL of 70% formic acid, air-dried, and overlaid with 1 μL of matrix solution. After crystallization, mass spectra were acquired and matched against the EXS2000 reference database. Isolates with low-confidence or ambiguous identifications were further analyzed by 16S rRNA gene sequencing.

Nearly full-length 16S rRNA genes were amplified using universal primers 27F and 1492R and subjected to bidirectional Sanger sequencing. Forward and reverse reads were assembled into consensus sequences and compared with reference sequences using NCBI BLAST. Sequence identity ≥ 98.65% with high query coverage was used as the reference criterion for species-level assignment, and ambiguous matches among closely related species were assigned conservatively. For isolates subjected to whole-genome sequencing (WGS), species identity was further evaluated by average nucleotide identity (ANI), with ≥ 95% used as the species-level threshold. All isolates were assigned unique strain codes and cryopreserved at − 80°C.

2.2 Whole-Genome Sequencing, Assembly, and Quality Control

A total of 172 strains were selected for WGS according to source diversity and genomic representativeness, including all 50 *L. fermentum* isolates, all 30 *L. plantarum* isolates, and 92 *L. paracasei* isolates selected by stratified random sampling from 839 recovered isolates. Strain metadata, Genome/WGS accession numbers, and genome assembly characteristics are provided in Table S1. Genomic DNA was extracted and quality controlled before library construction. DNA was randomly sheared, end-repaired, A-tailed, ligated with Illumina adapters, size selected, PCR amplified, and purified. Libraries were quantified using Qubit and real-time PCR, checked for fragment-size distribution using an Agilent Bioanalyzer, pooled, and sequenced on an Illumina platform (Novogene, Beijing, China) using paired-end sequencing.

Raw reads were quality filtered using fastp (v0.23.4) to remove adapters, reads containing > 10% ambiguous bases, and low-quality reads containing > 40% bases with Q ≤ 20. Read quality was checked using FastQC (v0.12.1). High-quality reads were assembled de novo using Unicycler (v0.5.1) (Wick et al. 2017). Assembly quality was evaluated using QUAST (v5.3.1) on the Galaxy platform, including total assembly length, GC content, contig number, and N50 (Mikheenko et al. 2018).

Genome annotation was performed using Prokka (Galaxy Version 1.14.6 + galaxy1) on the Galaxy platform (Seemann 2014). The resulting GFF files were used as input for Roary, and predicted protein sequences were used for eggNOG-mapper, InterProScan, and BLASTP-based analyses.

2.3 Pan-genome, Core-genome, and Phylogenetic Analysis

Orthologous core- and pan-genome analyses were performed separately for each species using Roary (v3.13.0) at a 95% sequence-identity threshold (Page et al. 2015). Gene clusters were categorized as core (≥ 95% of strains), shell (15% to < 95%), or cloud (< 15%) genes. Pan-genome profiles were

visualized in R (v4.3.3). Neighbor-joining phylogenetic trees were constructed in MEGA X from core-genome alignments with 1,000 bootstrap replicates. Pan-genome openness was evaluated using Heaps' law, P(N) = κN^γ, with fitting parameters estimated by nonlinear regression after 1,000 random permutations. Intraspecies genomic similarity was evaluated by ANI using dRep (v3.6.2) (Olm et al. 2017). Reference strains shown in the representative phylogeny and ANI panels (Fig. 2D–F) were included for visualization only and excluded from pan-genome statistics.

2.4 Genomic Functional Annotation and Comparative Analysis

Genome-based functional annotation was conducted using eggNOG-mapper (v2.1.8), from which Clusters of Orthologous Groups (COG), Kyoto Encyclopedia of Genes and Genomes (KEGG), and carbohydrate-active enzyme (CAZyme)-related annotations were extracted (Cantalapiedra et al. 2021). CAZyme-associated genes were classified into glycoside hydrolases (GHs), glycosyltransferases (GTs), carbohydrate esterases (CEs), carbohydrate-binding modules (CBMs), auxiliary activities (AAs), and polysaccharide lyases (PLs). Conserved protein domains and functional signatures were additionally annotated using InterProScan (Galaxy Version 5.59-91.0 + galaxy3) (Jones et al. 2014).

COG- and KEGG-based functional profiles were analyzed in R. Features with zero variance were removed before principal component analysis (PCA), which was performed using prcomp() with scaling. Bray–Curtis dissimilarities were calculated using vegdist(), and interspecies differences in overall functional composition were tested by permutational multivariate analysis of variance (PERMANOVA) using adonis2() in the vegan package. Functional annotations were integrated with Roary-derived pan-genome partitions to compare COG categories, KEGG pathways, and CAZyme-associated genes among core, shell, and cloud clusters.

InterProScan annotations were mapped to corresponding Roary orthologous clusters using protein identifiers. Repeated occurrences of the same InterPro feature within a single orthologous cluster were counted once. Selected functional genomic features were grouped into bile and lipid metabolism, acid and stress responses, adhesion and colonization, EPS biosynthesis, short-chain fatty-acid (SCFA) metabolism, amino acid metabolism, secondary metabolites, and antibacterial potential. Functional counts were normalized to 1,000 gene clusters. Species-level values were calculated as mean normalized counts and interpreted descriptively unless a specific inferential test was performed. For representative strains, each feature was independently standardized by Z-score for heatmap visualization.

Potential antimicrobial resistance and virulence-associated genes were screened against the Comprehensive Antibiotic Resistance Database (CARD) and Virulence Factor Database (VFDB), respectively, using ABRicate (v1.0.1). A uniform threshold of ≥ 80% sequence identity and ≥ 60% alignment coverage was applied. Comparable identity and coverage criteria have been used in previous genomic safety assessments of LAB (Hsu et al. 2022).

2.5 BSH Sequence and Structural Analysis

BSH homologs in *L. plantarum*, *L. fermentum*, and *L. paracasei* were identified by BLASTP against predicted proteomes (E-value < 0.001), supplemented by annotation-guided curation where assignments were incomplete or ambiguous. Final BSH copy numbers and protein sets were determined by integrating homology-search and annotation evidence. Identical sequences were collapsed into amino-acid haplotypes, and eight study-derived representatives together with seven reference BSHs were aligned using MAFFT (Katoh and Standley 2013). Maximum-likelihood phylogeny was inferred with IQ-TREE using the Q.pfam + F+I+R2 model and 1,000 ultrafast bootstrap replicates (Minh et al. 2020). Amino-acid substitutions were assessed relative to lineage-specific references, including conservation of catalytic Cys2.

Structural models were generated for the eight representative BSHs using ColabFold on Galaxy (n = 6) (Mirdita et al. 2022) or SWISS-MODEL with highly homologous AlphaFold DB templates (n = 2) (Varadi et al. 2022; Waterhouse et al. 2018). Ligand-bound structures 8BLS and 8BLT, containing glycocholate (GCH) and taurocholate (TCH), respectively, were used as references (Karlov et al. 2023). Structural comparisons were performed with Bio3D (Grant et al. 2006) and included structural-core root-mean-square deviation (RMSD), ligand-proximal RMSD, and the shortest Cys2-to-ligand distance. Ligand-proximal residues were defined within 6 Å for quantitative analysis and 4 Å for visualization; substitutions were considered pocket related when located at homologous ligand-pocket residues or within 8 Å of the ligand after local structural fitting.

2.6 EPS/CPS Locus Analysis

Exopolysaccharide/capsular polysaccharide (EPS/CPS) loci were characterized according to major regulatory, biosynthetic, and transport components, including regulators, priming GTs, bacterial tyrosine kinases (BY-kinases), kinase modulators, protein tyrosine phosphatases (PTPs), Wzx polysaccharide flippases, Wzy polysaccharide polymerases, Wzz polysaccharide chain-length determinants, Wzm ABC-transporter permeases, and Wzt ABC-transporter ATP-binding proteins. Component prevalence was calculated separately for each species.

To reduce redundancy in locus visualization, EPS/CPS regions were compared within each species using weighted Jaccard similarity based on functional-component counts. Regions were considered highly similar when similarity was ≥ 0.90 and the locus-length ratio ranged from 0.80 to 1.25. One representative region was retained from each highly similar group. Gene positions were expressed relative to the start of each retained locus. Because the genomes were draft assemblies, reconstructed locus length was treated as a descriptive property of the assembled region and was not assumed to represent complete biological locus size when contig fragmentation could affect locus boundaries.

2.7 Determination of Simulated Gastrointestinal Transit Tolerance

Fourteen study-derived strains representing the phylogenetic and genomic diversity of the three focal species were selected for phenotypic evaluation: *L. plantarum* R2345, R203, R1621, R3644, and R2037; *L. paracasei* R1156, B1023, R11515, R2670, B408, and B11517; and *L. fermentum* B402, B1040, and B4036. *Lacticaseibacillus rhamnosus* GG (LGG), maintained in our laboratory culture collection, was used as the reference strain for phenotypic assays.

Strain tolerance to simulated gastrointestinal conditions was evaluated using commercially prepared simulated gastric fluid (SGF, pH 3.0) and simulated intestinal fluid (SIF, pH 7.4), both purchased from Shanghai Yuanye Bio-Technology Co., Ltd. (Shanghai, China). Activated cultures were mixed with SGF at a 1:10 (v/v) ratio and incubated anaerobically at 37°C for 2 h. Suspensions were then transferred to SIF at the same ratio and incubated for 3 h under identical conditions. Samples were collected at each stage for viable-cell enumeration (CFU/mL) by standard plate counting. Three independently prepared culture replicates were analyzed for each strain.

Survival rate (%) = $(N_1/N_0) \times 100$

where N_0 is the initial viable count and N_1 is the viable count after exposure.

2.8 Evaluation of In Vitro Cholesterol-Removal Capacity

Cholesterol-removal capacity was evaluated under three culture conditions: standard MRS, 10-fold diluted MRS, and MRS supplemented with 0.1% (w/v) bovine bile salts (Solarbio, Beijing, China). The 10-fold diluted MRS condition was used as a standardized nutrient-reduced environment to assess whether cholesterol-removal capacity was retained under reduced nutrient availability. For the bile-supplemented condition, bovine bile salts were added to standard MRS to a final concentration of 0.1% (w/v) to assess cholesterol removal in the presence of bile salts. Each medium contained 0.1 mg/mL cholesterol. Activated cultures were adjusted to OD_{600} = 1.0, inoculated at 1% (v/v), and cultured anaerobically at 37°C to logarithmic phase.

Cultures were centrifuged at 8,000 × g and 4°C for 8 min. Supernatants were filtered through 0.22 μm membranes, and residual cholesterol was quantified using a commercial cholesterol oxidase–peroxidase assay kit (Beyotime, Shanghai, China) at 510 nm.

Cholesterol removal (%) = $(C_0 - C_1)/C_0 \times 100$

where C_0 is the cholesterol concentration in the uninoculated blank and C_1 is the residual cholesterol concentration in the filtered supernatant.

2.9 Statistical Analysis

Quantitative data are presented as mean ± SD from three independently prepared culture replicates (n = 3) for each strain and condition; each culture replicate was measured once. For simulated gastrointestinal tolerance, gastric and intestinal endpoints were analyzed separately by one-way analysis of variance (ANOVA) followed by Dunnett’s multiple-comparison test with LGG as the reference.

Cholesterol-removal data were analyzed by two-way ANOVA with strain, culture condition, and their interaction as fixed effects. Following a significant strain × culture-condition interaction, simple effects were examined using estimated marginal means. Within each culture condition, strains were compared with LGG using Dunnett-adjusted contrasts, whereas culture conditions were compared within each strain using Tukey-adjusted pairwise contrasts. Multiplicity-adjusted P values < 0.05 were considered statistically significant.

3. Results

3.1 Isolation and Taxonomic Identification of Sour Soup-Derived Strains

A total of 1,741 culturable microbial isolates were recovered from the eight sour soup samples. Lactobacillales accounted for 85.69% and 95.24% of isolates recovered from red and white sour soups, respectively (Fig. 1A, B), whereas Eubacteriales accounted for 11.61% and 1.68%, respectively.

At the genus level, *Lacticaseibacillus*, *Schleiferilactobacillus*, *Lactobacillus*, and *Lentilactobacillus* were frequently recovered (Fig. 1C). Red and white sour soups yielded 35 and 13 unique culturable species, respectively, with 12 species shared between them (Fig. 1D).

At the species level, recovery patterns differed between the two sour-soup types (Fig. 1E). *L. paracasei* was recovered more frequently from white sour soup (615 isolates) than from red sour soup (224 isolates). *L. buchneri* was represented by 112 red-soup and 21 white-soup isolates, whereas *L. brevis* was represented by 68 and 8 isolates, respectively. *L. fermentum* was recovered from B1 (11 isolates), B4 (26), and R1 (13), whereas *L. plantarum* was recovered from R1 (18), R2 (10), and R3 (2).

3.2 Whole-Genome Assembly and Population Genetic Analysis

Whole-genome sequencing of 172 isolates (92 *L. paracasei*, 30 *L. plantarum*, and 50 *L. fermentum*) yielded draft genome assemblies ranging from 1.79 to 3.38 Mb in total length and from 44.29% to 52.77% in GC content (Table S1). *L. fermentum* had the smallest mean assembly size and the highest mean GC content among the three species.

Heaps' law analysis ($0 < \gamma < 1$, $P < 0.001$) supported open pan-genomes for all three species (Fig. 2A; Table 1). Pan-genome composition differed among species (Fig. 2B): cloud genes accounted for 48.7% of the *L. paracasei* pan-genome, whereas core genes accounted for 62.0% and 60.2% of the *L. plantarum* and *L. fermentum* pan-genomes, respectively. Because the three species were represented by different numbers of genomes, absolute cluster counts were treated as dataset-specific descriptors rather than as a direct ranking of genomic plasticity.

Gene presence/absence clustering, core-genome phylogeny, and ANI analysis showed species- and strain-level structure (Fig. 2C–F). ANI-based clustering resolved *L. paracasei* into nine secondary clusters (ANI > 98.40%), whereas *L. plantarum* and *L. fermentum* each formed three clusters (ANI > 99.00% and > 99.20%, respectively).

Table 1
Pan-genome composition and Heaps' law parameters of the three focal LAB species.

Species	No. of genomes	Core clusters	Shell clusters	Cloud clusters	Cloud proportion (%)	Heaps' γ
L. paracasei	92	1,827	1,965	3,605	48.7	0.1797
L. plantarum	30	2,490	994	534	13.3	0.0756
L. fermentum	50	1,451	646	315	13.1	0.1188

Note. Values are based on Roary gene-cluster partitions generated separately for each species. Heaps' law supported open pan-genomes for all three species ($0 < \gamma < 1$, $P < 0.001$).

3.3 Genomic Functional Annotation Profiles

3.3.1 Functional Diversity and Pan-genome Distribution

COG- and KEGG-based functional profiles separated the three species (Fig. 3A, B). In the COG PCA, PC1 and PC2 explained 84.1% and 10.1% of total variation, respectively; PERMANOVA supported interspecies differences in overall COG composition ($R^2 = 0.927$, $P = 0.001$). KEGG PCA showed a similar pattern, with PC1 and PC2 explaining 60.6% and 21.9% of variation (PERMANOVA $R^2 = 0.795$, $P = 0.001$). Intraspecies PCA showed additional subgroup-level structure.

Mapping functional annotations to pan-genome partitions showed different distributions of conserved and accessory functions among species (Fig. 3C). Core genes accounted for a relatively large proportion of major COG categories and KEGG pathways in *L. plantarum* and *L. fermentum*. In *L. paracasei*, shell and cloud genes contributed more broadly to metabolic, transport, regulatory, and other cellular functions.

CAZyme-associated clusters followed a similar descriptive pattern (Fig. 3D). GHs and GTs were the most represented CAZyme classes in all three species. In *L. plantarum* and *L. fermentum*, these functions were more frequently assigned to core clusters, whereas *L. paracasei* showed a larger shell/cloud contribution. Similar patterns were observed for CE, CBM, AA, and PL classes.

3.3.2 Selected Functional Genomic Profiles

Selected functional genomic profiles differed descriptively among species after normalization to 1,000 gene clusters (Fig. 4). *L. paracasei* showed relatively high representation of several stress-response and cell-surface-associated features, including thioredoxin-, adhesion-, and colonization-associated

functions. CoA transferase and several antibacterial-associated features were also represented in this species.

L. fermentum showed relatively high representation of F-type ATPase, NADH oxidase, peroxiredoxin/AhpC, amino-acid transport, and aminotransferase-associated features. *L. plantarum* had the highest mean normalized BSH representation and retained multiple EPS-, SCFA-, and stress-associated functions.

Feature-wise Z-score standardization revealed substantial variation among representative strains (Fig. 4). Adhesion and colonization, EPS biosynthesis, SCFA metabolism, and antibacterial-associated functions showed pronounced strain-level heterogeneity. Because each feature was standardized independently, the heatmap represents relative within-feature variation rather than absolute differences among functional categories.

3.3.3 BSH Diversity and Structural Conservation

BSH phylogeny revealed species- and lineage-specific patterns (Fig. 5A). *L. plantarum* strains carried 3–4 BSH copies representing four lineages, with R3644 and R3632 lacking the BSH3536-like copy. *L. fermentum* strains consistently carried two BSH lineages, whereas *L. paracasei* carried 1–2 BSH0412-like copies; H02 was present in all 92 strains and H01 occurred additionally in 11 strains. Representative BSHs showed lineage-specific sequence variation, ranging from 0–5 substitutions in *L. plantarum* and 1–2 in *L. fermentum*, while the two *L. paracasei* haplotypes contained 14–17 substitutions. Catalytic Cys2 was conserved in all representatives, and no substitutions met the predefined pocket-related criterion.

Structural comparisons showed overall conservation of the catalytic framework (Fig. 5C, D). Structural-core RMSD ranged from 0.961 to 1.597 Å and ligand-proximal RMSD from 0.662 to 1.607 Å; Cys2-to-GCH and Cys2-to-TCH distances were 3.364–3.898 Å and 5.055–5.799 Å, respectively. These comparisons support structural conservation among the representative BSHs without implying substrate preference or equivalent catalytic activity.

3.3.4 EPS/CPS Locus Diversity

EPS/CPS component prevalence differed among the three species (Fig. 5B). *L. plantarum* showed the most conserved profile, with priming GT, BY-kinase, kinase modulator, PTP, and Wzz present in all strains. *L. fermentum* also retained a recurrent core set: regulators, BY-kinase, kinase modulators, and PTP were present in 96% of strains, Wzz in 100%, and priming GT in 50%. *L. paracasei* showed greater variation, with BY-kinase and Wzz detected in 90% and 77% of strains, respectively, whereas kinase modulators (41%), priming GT (28%), Wzy (10%), regulators (6%), PTP (6%), and Wzm (12%) were less frequent.

EPS/CPS locus organization also varied among strains (Fig. 5E). In *L. plantarum*, R2037 and R2345 showed similar locus architectures, while R203, R1621, and R1610 formed another similar group. R3644 and R3632 showed shorter reconstructed EPS/CPS regions with distinct gene organizations. *L. paracasei* showed greater variation in reconstructed region length, gene order, and glycosyltransferase

composition. B2351, B4047, and B41211 shared similar architectures, as did R2670 and R1156, and B3035 and B408. B2312, B3035, and B408 showed relatively shorter reconstructed regions, whereas B1634, B2351, B4047, and B41211 showed longer regions. In *L. fermentum*, B1040, R1337, B108, B4036, and B41219 showed relatively extended reconstructed regions, while B402, B463, and B4154 showed shorter regions with similar organizations.

3.4 Genome-Based Safety Screening

Whole-genome safety screening detected no antimicrobial resistance or virulence-associated genes meeting the applied thresholds (≥ 80% sequence identity and ≥ 60% alignment coverage) in any of the 172 genomes.

3.5 Targeted In Vitro Phenotypic Evaluation

To determine whether the genomic heterogeneity described above translated into measurable functional differences, representative strains spanning the phylogenetic and genomic diversity of the three focal species were evaluated for simulated gastrointestinal tolerance and cholesterol-removal capacity.

3.5.1 Survival Rates in Simulated Gastrointestinal Transit

Simulated gastrointestinal tolerance varied among strains (Fig. 6A, B). After gastric exposure, *L. plantarum* R2345 showed 95.62 ± 2.16% survival, followed by *L. paracasei* R1156 (93.07 ± 3.20%), *L. plantarum* R203 (92.61 ± 1.75%), and *L. paracasei* B1023 (92.61 ± 1.75%); LGG showed 86.74 ± 2.33%. Most strains did not differ from LGG, whereas *L. plantarum* R3644 showed lower survival (54.16 ± 3.57%, adjusted $P < 0.0001$).

After intestinal exposure, *L. plantarum* R2345 retained 93.07 ± 2.69% survival. *L. paracasei* B1023 and *L. plantarum* R203 each showed 87.19 ± 6.44%, compared with 85.97 ± 1.57% for LGG. *L. fermentum* B402 and *L. plantarum* R3644 showed lower survival (46.07 ± 4.29% and 45.87 ± 5.54%, respectively; both adjusted $P < 0.0001$).

3.5.2 Cholesterol-Removal Capacity

Cholesterol-removal performance showed marked strain- and condition-dependent variation across the three culture conditions (Fig. 6C–E; Table S3A). Two-way ANOVA revealed a significant strain × culture-condition interaction ($F(28, 90) = 39.99$, $P < 0.0001$; Table S3C).

In standard MRS (Fig. 6C), *L. paracasei* R1156 and B11517 showed the highest removal rates, reaching 71.69 ± 4.84% and 69.44 ± 4.43%, respectively. *L. plantarum* R2345 showed 53.03 ± 2.59% removal, compared with 37.78 ± 0.98% for LGG.

Under 10-fold diluted MRS (Fig. 6D), *L. plantarum* R2037, R2345, and R3644 reached 94.04 ± 0.40%, 93.28 ± 2.74%, and 90.92 ± 1.54%, respectively, compared with 88.99 ± 0.71% for LGG; none showed a significant increase relative to LGG after multiplicity adjustment.

Under MRS supplemented with 0.1% bile salts (Fig. 6E), *L. plantarum* R2345, R2037, R1621, and R203 maintained removal rates of 64.32 ± 0.93%, 63.15 ± 0.54%, 58.24 ± 1.70%, and 56.85 ± 0.29%, respectively, all significantly higher than LGG (23.91 ± 0.26%; adjusted $P < 0.0001$). *L. paracasei* B408 also exceeded LGG (36.58 ± 2.68%, adjusted $P = 0.0006$). Pairwise comparisons among culture conditions within each strain are provided in Table S3B.

4. Discussion

The principal finding was that the three co-occurring LAB species exhibited distinct patterns in pan-genome architecture and functional gene distribution, while considerable within-species genomic variation did not translate into a simple genomic prediction of phenotype. Red and white Guizhou sour soups yielded different culturable LAB profiles, consistent with previous reports of distinct bacterial communities between the two fermentation types (Zhou et al. 2026). Amplicon sequencing and enrichment-based culturomics capture different fractions of the microbiota; the two approaches are therefore complementary rather than directly comparable. All three focal species exhibited open pan-genomes with species-specific distributions of core and accessory genes, consistent with comparative genomic studies showing substantial within-species functional differentiation among food-associated lactobacilli (Mejía-Caballero et al. 2025). Together, these patterns indicate that Guizhou sour soup harbors a heterogeneous living LAB resource in which species-level genomic architecture and strain-level accessory variation can be evaluated within a shared fermented-food context.

The relationship between genomic features and phenotype was not straightforward. EPS/CPS loci differed in composition and organization among strains, but these differences did not show a simple correspondence with simulated gastrointestinal survival. *L. plantarum* strains with different reconstructed EPS/CPS architectures could display similarly high survival, whereas R3644, with a distinct architecture, showed much lower tolerance. The observed phenotype is likely to involve multiple stress-response and cell-surface-associated factors rather than EPS/CPS architecture alone. This lack of a simple one-to-one correspondence suggests that EPS/CPS loci are best treated as one component of a broader genome-informed screening framework rather than as stand-alone predictors of functional performance. Direct measurements of EPS production, composition, and expression will be needed to determine whether the genomic differences observed here contribute to gastrointestinal tolerance.

BSH diversity provided a second example of the limits of genotype-based prediction. Although representative BSH proteins retained catalytic Cys2 and a conserved ligand-proximal architecture, BSH copy number and lineage composition alone did not account for cholesterol-removal performance. *L. plantarum* R3644, which lacked the BSH3536-like lineage, showed high cholesterol removal in 10-fold diluted MRS but lower removal under bile-containing conditions; several other *L. plantarum* strains maintained higher removal in the presence of bile. Two-copy *L. paracasei* representatives showed no consistent advantage over single-copy strains under bile-containing conditions. These observations are consistent with previous evidence that bile-salt metabolism and cholesterol removal are not necessarily coupled (Zhang et al. 2025). BSH therefore provides useful genomic context for bile-associated

phenotypes but is not a stand-alone predictor of cholesterol removal, reinforcing the need to combine genome-based screening with direct functional assays when prioritizing strains.

The significant strain × culture-condition interaction further emphasizes that cholesterol-removal capacity is context dependent. Several *L. plantarum* strains showed removal rates above 90% in 10-fold diluted MRS, yet none was significantly superior to LGG after multiple-comparison adjustment, and their relative rankings changed under bile-containing conditions. A high removal rate measured in a single medium should not be treated as an intrinsic and stable property of a strain. This has practical implications for functional screening because strain rankings obtained under one assay condition may not remain stable when nutrient or bile conditions change. Evaluation across multiple relevant conditions therefore provides a more robust basis for distinguishing functional profiles and prioritizing candidates for subsequent application-oriented testing.

Taken together, strain-level genomic diversity and phenotypic diversity were related but not interchangeable. Comparative genomics can define genetic backgrounds and prioritize informative isolates, but direct phenotypic testing remains necessary to establish functional performance. This distinction is particularly important for traditional fermented foods, where closely related isolates may carry substantial accessory-genome variation and still differ markedly in measurable phenotype. Strain-resolved genomics is therefore most useful when it guides functional screening rather than substitutes for it. An integrated genome-informed and phenotype-verified strategy provides a more rational basis for selecting food-associated LAB for downstream biotechnological evaluation.

Genome-based screening identified no antimicrobial resistance or virulence-associated genes at the applied thresholds, providing preliminary genomic evidence relevant to strain selection; however, genomic screening does not replace phenotypic antimicrobial susceptibility testing or broader safety assessment. Several isolates emerged as candidates for further validation. *L. plantarum* R2345 combined favorable simulated gastrointestinal tolerance with relatively high cholesterol removal under bile-containing conditions, whereas R203, R2037, and R1621 showed distinct complementary phenotypic profiles. *L. paracasei* R1156 and B11517 showed high cholesterol removal in standard MRS. These strains should be regarded as candidates rather than strains with established health benefits, and their functional value will require further mechanistic and in vivo validation. They nevertheless constitute a focused set for subsequent process-relevant and application-oriented testing. By linking a living culture collection to genome-resolved and phenotype-resolved data, this study provides a practical resource for genome-informed selection of food-associated LAB while showing that direct phenotypic confirmation remains essential.

5. Conclusion

This study established a large-scale culturable LAB resource from red and white Guizhou sour soups and integrated comparative pan-genomics with targeted in vitro phenotypic evaluation to link strain-level genomic diversity with measurable functional performance.

Red and white sour soups yielded distinct culturable LAB profiles. All three focal species had open pan-genomes but differed in genomic architecture and functional representation. *L. paracasei* showed a larger cloud-gene fraction and broader accessory-genome contribution, *L. plantarum* contained a broader BSH repertoire together with multiple EPS- and stress-associated features, and *L. fermentum* showed distinct oxidative-stress- and amino-acid-associated profiles. These differences define complementary genomic backgrounds for genome-informed functional screening.

Genome-based safety screening detected no antimicrobial resistance or virulence-associated genes meeting the applied thresholds in any of the 172 genomes, providing preliminary genomic evidence relevant to subsequent strain prioritization.

Targeted phenotypic assays revealed clear strain- and condition-dependent variation. *L. paracasei* R1156 and B11517 showed high cholesterol removal in standard MRS, whereas *L. plantarum* R2345, R2037, R1621, and R203 maintained 56.85–64.32% removal under 0.1% bile salts. Several strains also retained high removal in 10-fold diluted MRS. The marked strain × condition effect, together with the lack of a simple correspondence between BSH features and cholesterol-removal performance, demonstrates that genomic traits alone are insufficient to predict functional performance.

Among the representative strains, *L. plantarum* R2345 and R203 combined favorable gastrointestinal tolerance with relatively high cholesterol removal under bile-containing conditions, while R2037 and R1621 showed strong bile-conditioned cholesterol removal. These strains therefore constitute a focused set for further mechanistic, process-relevant, and application-oriented validation rather than evidence of established health benefits.

Overall, this study reveals substantial strain-level genomic and phenotypic heterogeneity among culturable LAB from Guizhou sour soup and demonstrates the practical value of integrating comparative genomics with direct phenotypic screening. The resulting culture and genome resource provides a rational framework for prioritizing indigenous food-associated LAB for further functional and biotechnological evaluation.

Declarations

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declarations

of generative AI and AI-assisted technologies in the manuscript preparation process

Funding

This work was supported by the Opening Foundation of the National Health Commission (NHC) Key Laboratory of Etiology and Epidemiology (Harbin Medical University) and the Open Project of the Joint Key Laboratory of Endemic Diseases (Harbin Medical University, Guizhou Medical University, Xi'an Jiaotong University) [grant number NHCKLEE202504]. The funding sources had no involvement in the study design; collection, analysis, or interpretation of data; preparation of the manuscript; or the decision to submit the article for publication.

Author Contribution

FB: Formal analysis, Investigation, Writing – original draft. NW: Investigation. YL: Data curation. QW: Data curation. CH: Formal analysis, Data curation. SZ: Funding acquisition. RM: Supervision. RY: Methodology, Supervision, Writing – review & editing. YH: Supervision, Writing – review & editing. PL: Conceptualization, Project administration, Supervision, Writing – review & editing.

Data Availability

Genome assemblies generated in this study have been submitted to NCBI GenBank under BioProject accession PRJNA1491403. Individual accession numbers for fully processed strains are provided in Table S1. Remaining accessions currently in processing will be fully assigned and publicly accessible prior to formal publication.

References

1. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J (2021) eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol 38:5825–5829. https://doi.org/10.1093/molbev/msab293
2. Carlino N, Blanco-Míguez A, Punčochář M, Mengoni C, Pinto F, Tatti A et al (2024) Unexplored microbial diversity from 2,500 food metagenomes and links with the human microbiome. Cell 187:5775–5795e15. https://doi.org/10.1016/j.cell.2024.07.039
3. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD (2006) Bio3d: An R package for the comparative analysis of protein structures. Bioinformatics 22:2695–2696. https://doi.org/10.1093/bioinformatics/btl461
4. Hsu H-Y, Liao Y-C, Lin S-H, Lin J-S, Lee C-C, Watanabe K (2022) Safety assessment of *Lactiplantibacillus plantarum* TWK10 based on whole-genome sequencing, phenotypic, and oral toxicity analysis. Microorganisms 10:784. https://doi.org/10.3390/microorganisms10040784
5. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C et al (2014) InterProScan 5: Genome-scale protein function classification. Bioinformatics 30:1236–1240.

https://doi.org/10.1093/bioinformatics/btu031

6. Kahraman Ilıkkan Ö (2025) Lactobacillomics as a new notion in lactic acid bacteria research through omics integration. World J Microbiol Biotechnol 41:68. https://doi.org/10.1007/s11274-025-04285-y
7. Karlov DS, Long SL, Zeng X, Xu F, Lal K, Cao L, Hayoun K, Lin J, Joyce SA, Tikhonova IG (2023) Characterization of the mechanism of bile salt hydrolase substrate specificity by experimental and computational analyses. Structure 31:629–638e5. https://doi.org/10.1016/j.str.2023.02.014
8. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010
9. Marco ML, Sanders ME, Gänzle M, Arrieta MC, Cotter PD, De Vuyst L et al (2021) The International Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on fermented foods. Nat Rev Gastroenterol Hepatol 18:196–208. https://doi.org/10.1038/s41575-020-00390-5
10. Mejía-Caballero A, López-Sánchez R, Ramos-Cerrillo B, Garciarrubio A, Segovia L (2025) Genomic insights into habitat adaptation of Lactobacillus species. World J Microbiol Biotechnol 41:61. https://doi.org/10.1007/s11274-025-04275-0
11. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A (2018) Versatile genome assembly evaluation with QUAST-LG. Bioinformatics 34:i142–i150. https://doi.org/10.1093/bioinformatics/bty266
12. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R (2020) IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa015
13. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M (2022) ColabFold: Making protein folding accessible to all. Nat Methods 19:679–682. https://doi.org/10.1038/s41592-022-01488-1
14. Mukherjee A, Breselge S, Dimidi E, Marco ML, Cotter PD (2024) Fermented foods and gastrointestinal health: Underlying mechanisms. Nat Rev Gastroenterol Hepatol 21:248–266. https://doi.org/10.1038/s41575-023-00869-x
15. Olm MR, Brown CT, Brooks B, Banfield JF (2017) dRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J 11:2864–2868. https://doi.org/10.1038/ismej.2017.126
16. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG et al (2015) Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693. https://doi.org/10.1093/bioinformatics/btv421
17. Qayyum N, Wang S, Qin Y, Wang R, Wang S, Ismael M, Lü X (2023) Characterization of short-chain fatty acid-producing and cholesterol assimilation potential probiotic lactic acid bacteria from Chinese fermented rice. Food Biosci 52:102404. https://doi.org/10.1016/j.fbio.2023.102404
18. Seemann T (2014) Prokka: Rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153

19. Tian H, Huang N, Chen C, Yu H, Ge C (2025) Flavor profiles and genetic basis differences of *Lacticaseibacillus paracasei* isolates from different isolations in fermented milk. Int J Food Microbiol 440:111274. https://doi.org/10.1016/j.ijfoodmicro.2025.111274
20. Tjandrawinata RR, Kartawijaya M, Hartanti AW (2022) In vitro evaluation of the anti-hypercholesterolemic effect of Lactobacillus isolates from various sources. Front Microbiol 13:825251. https://doi.org/10.3389/fmicb.2022.825251
21. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G et al (2022) AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50:D439–D444. https://doi.org/10.1093/nar/gkab1061
22. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R et al (2018) SWISS-MODEL: Homology modelling of protein structures and complexes. Nucleic Acids Res 46:W296–W303. https://doi.org/10.1093/nar/gky427
23. Wick RR, Judd LM, Gorrie CL, Holt KE (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595
24. Zhang X, Zhao S, Cai S, Yi J, Wang T, Liu Z, Wang Y (2025) Screening cholesterol-lowering *Lactiplantibacillus plantarum* through in vitro experiments, whole genomics analysis, and in silico approaches: From the perspective of bile salt hydrolase and cholesterol. J Agric Food Chem 73:20779–20791. https://doi.org/10.1021/acs.jafc.5c03853
25. Zhou Q, Zhao C, Li X, Wang H, Huang Q, Sun Y, Zhou Y (2025) Discovery of lactic acid bacteria with high nucleoside degradation and low purine production in tomato sour soup. Int J Food Microbiol 434:111139. https://doi.org/10.1016/j.ijfoodmicro.2025.111139
26. Zhou R, Zhang W, Chen K, Lin Z, Yi X, Pan J, Chen X, Yang A (2026) Bacterial community and functional prediction differences between white and red sour soup. AMB Express Adv online publication. https://doi.org/10.1186/s13568-026-02080-y

Figures

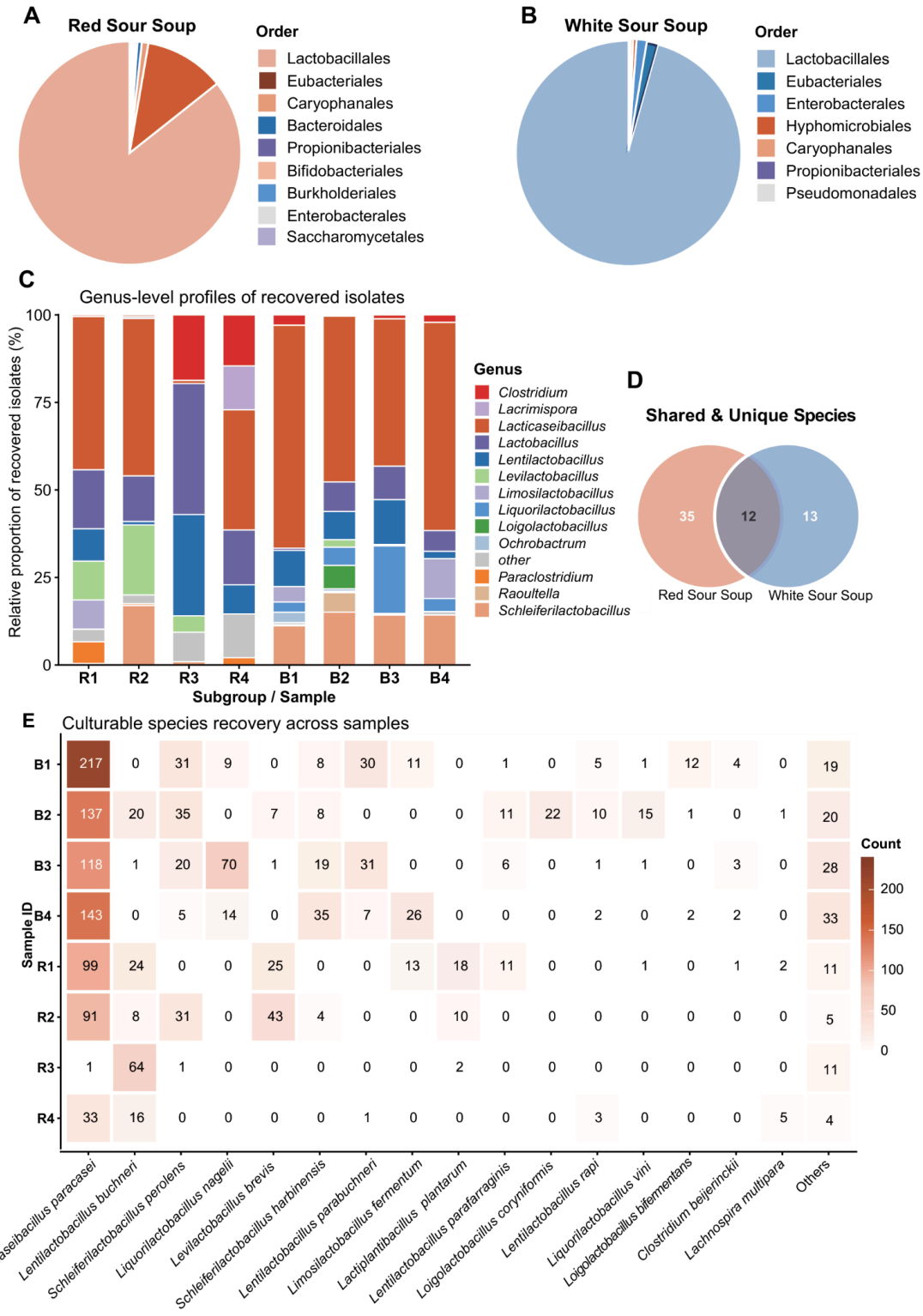


Figure 1

Culturable microbial profiles of Guizhou red and white sour soups. (A, B) Order-level composition of isolates recovered from red sour soup (A) and white sour soup (B). (C) Relative proportions of recovered genera across R1–R4 and B1–B4. (D) Shared and unique culturable species between the two sour-soup types. (E) Heatmap of isolate counts for representative species across samples. Values represent isolates recovered under the culturomics workflow

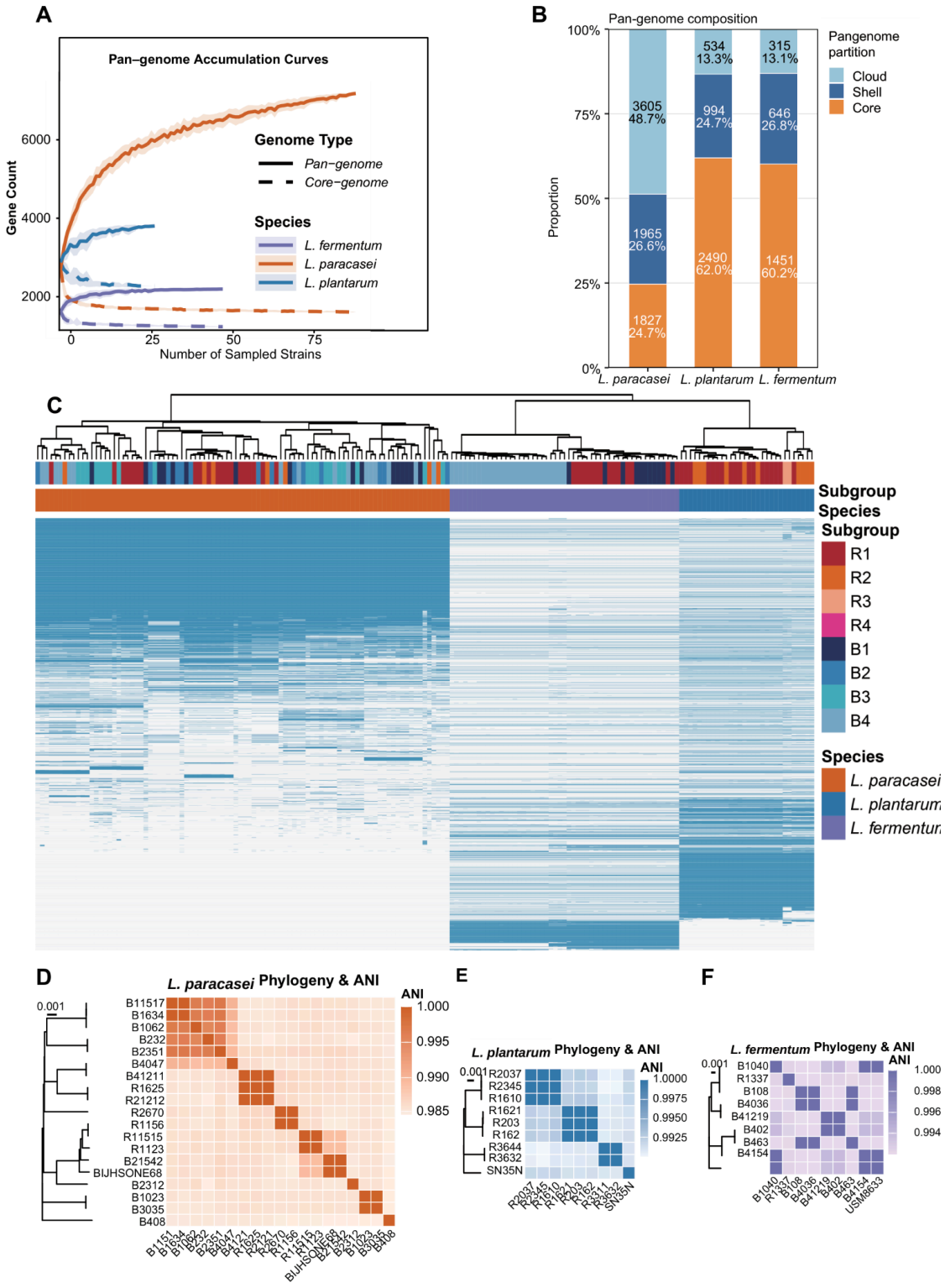


Figure 2

Pan-genome structure and genomic relationships of the three focal LAB species. (A) Pan-genome and core-genome accumulation curves for *L. paracasei*, *L. plantarum*, and *L. fermentum*. (B) Proportions and numbers of core, shell, and cloud gene clusters. (C) Hierarchical clustering of gene presence/absence profiles, with species and source subgroup annotations. (D–F) Core-genome phylogenies integrated with pairwise ANI matrices for representative *L. paracasei* (D), *L. plantarum* (E), and *L. fermentum*(F) strains.

Complete phylogenetic trees are provided in Figs. S1–S3, and complete pairwise ANI matrices are provided in Table S2

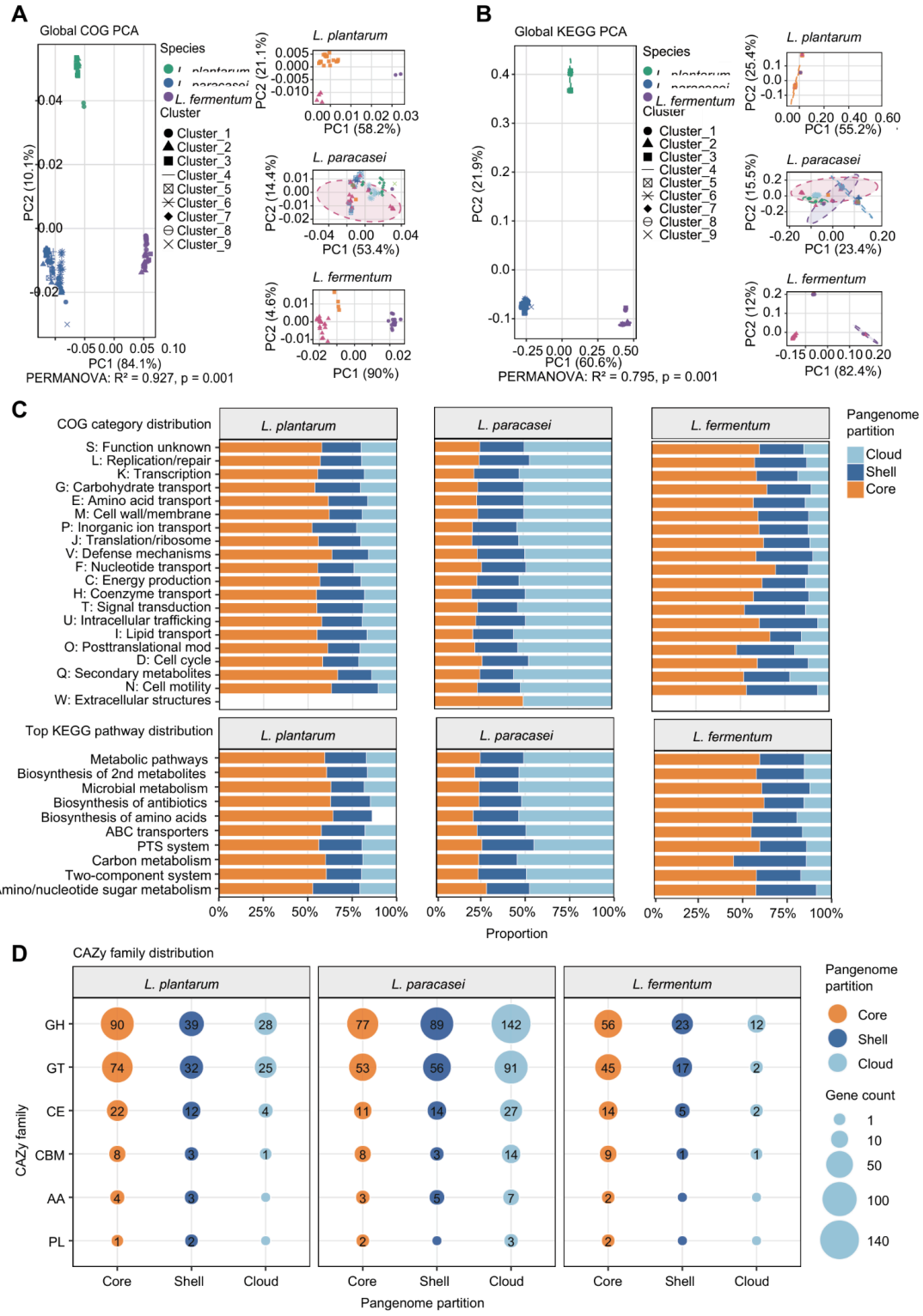


Figure 3

Comparative functional profiles and pan-genome partitioning. (A) Global COG-based PCA with species-specific insets; interspecies separation was evaluated by PERMANOVA. (B) Global KEGG-based PCA with

species-specific insets and PERMANOVA. (C) Distribution of COG categories and selected KEGG pathways across core, shell, and cloud partitions for each species. (D) Distribution of CAZyme-associated gene clusters across pan-genome partitions. Bubble size represents gene-cluster count. GH, glycoside hydrolase; GT, glycosyltransferase; CE, carbohydrate esterase; CBM, carbohydrate-binding module; AA, auxiliary activity; PL, polysaccharide lyase

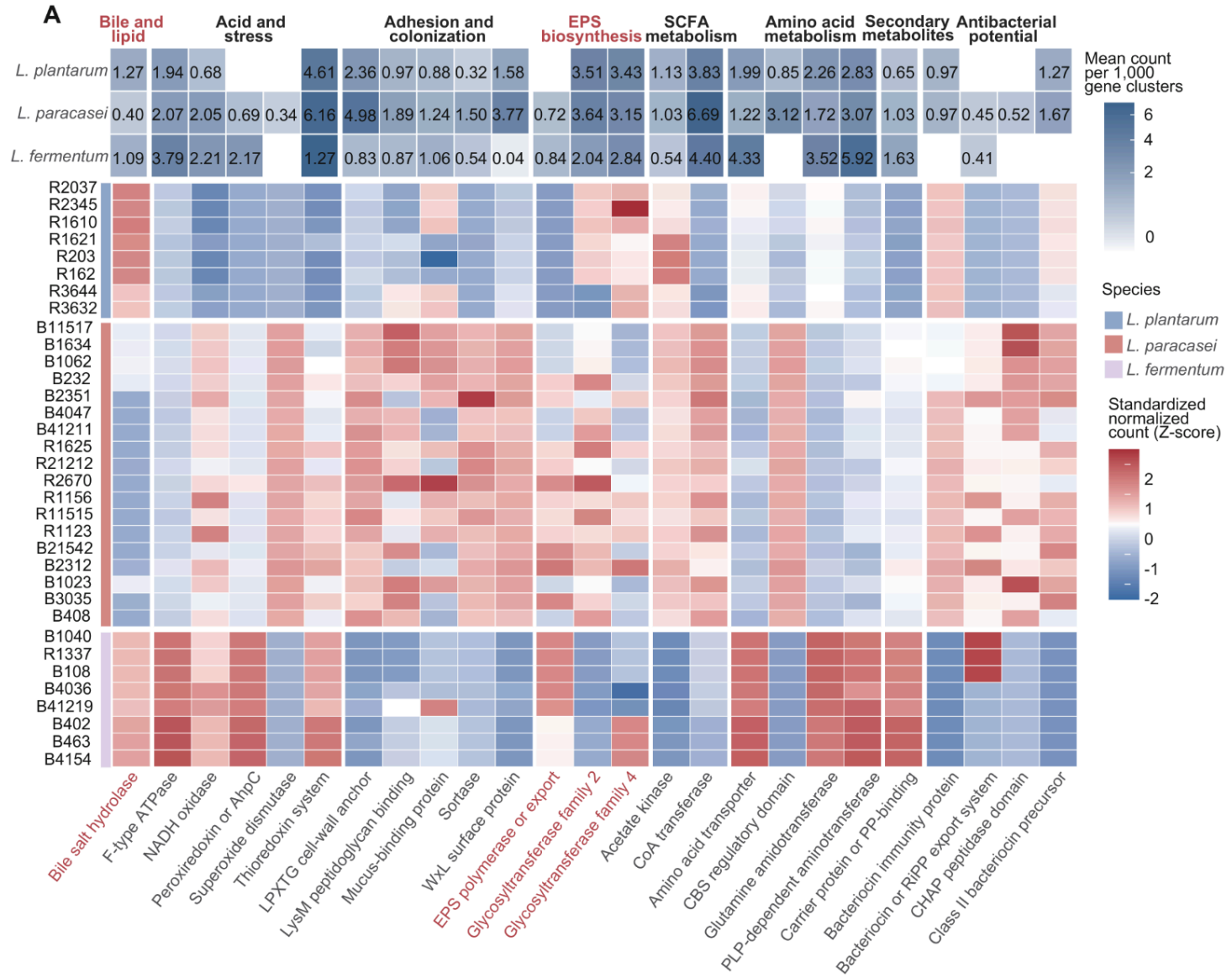


Figure 4

Selected functional genomic profiles of representative sour soup-derived LAB. Species-level mean normalized counts per 1,000 gene clusters are shown in the upper matrix. The lower heatmap shows feature-wise Z-score-standardized values for representative strains. Features are grouped into bile and lipid metabolism, acid and stress responses, adhesion and colonization, EPS biosynthesis, SCFA metabolism, amino-acid metabolism, secondary metabolites, and antibacterial potential. Species-level feature counts are descriptive

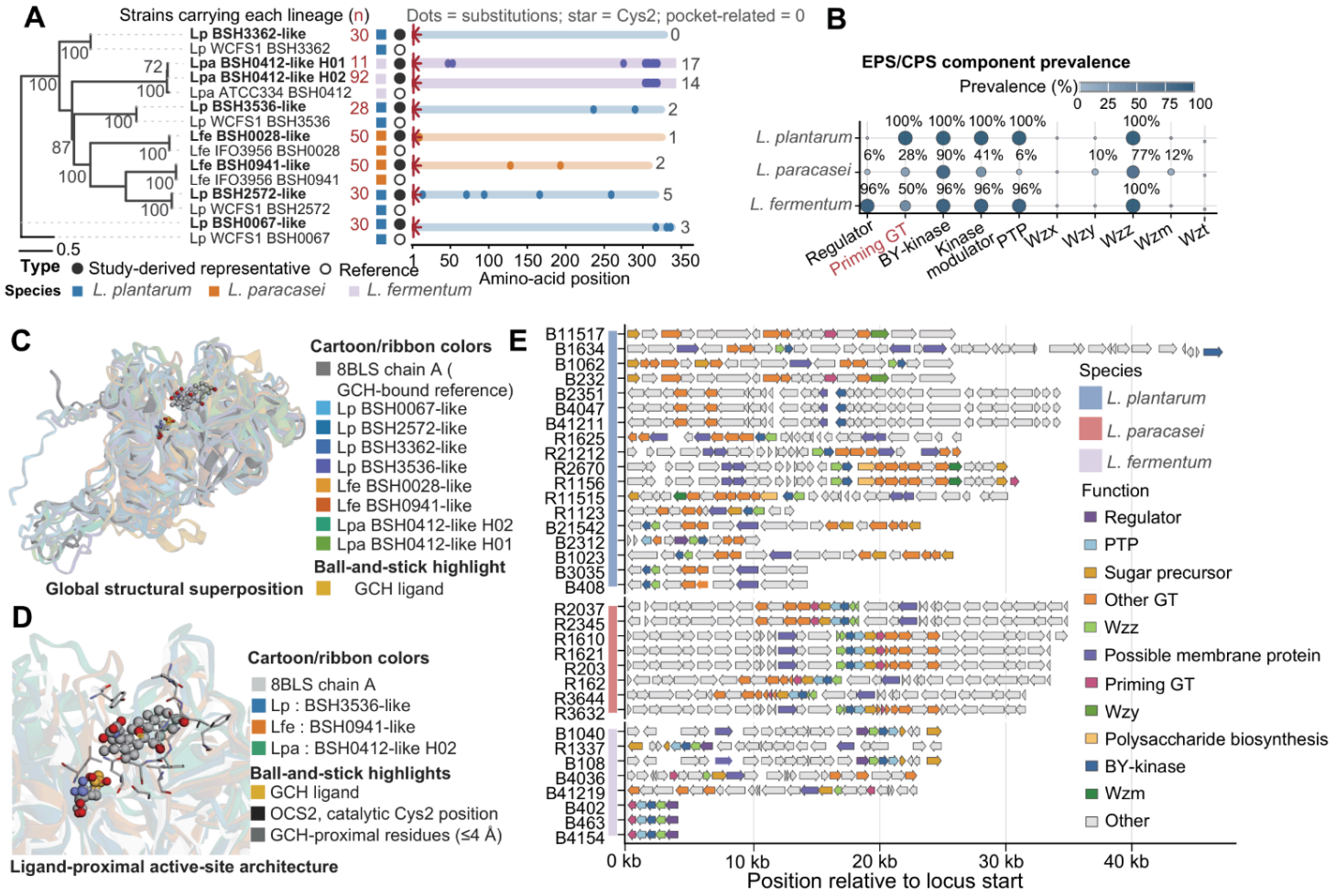


Figure 5

BSH and EPS/CPS genomic features of representative LAB strains. (A) Maximum-likelihood BSH phylogeny of seven reference BSHs and eight representative study-derived BSHs with lineage-specific amino-acid substitution maps. Numbers indicate strains carrying each lineage; all 92 *L. paracasei* strains carried H02, and 11 of these additionally carried H01. The catalytic Cys2 position is indicated, and no representative substitution met the predefined pocket-related criterion. (B) Prevalence of major EPS/CPS components across species. GT, glycosyltransferase; BY-kinase, bacterial tyrosine kinase; PTP, protein tyrosine phosphatase. (C) Global structural superposition of representative BSH models with the GCH-bound experimental reference. (D) Ligand-proximal active-site architecture highlighting catalytic Cys2 and GCH-proximal residues. (E) Organization of representative reconstructed EPS/CPS regions

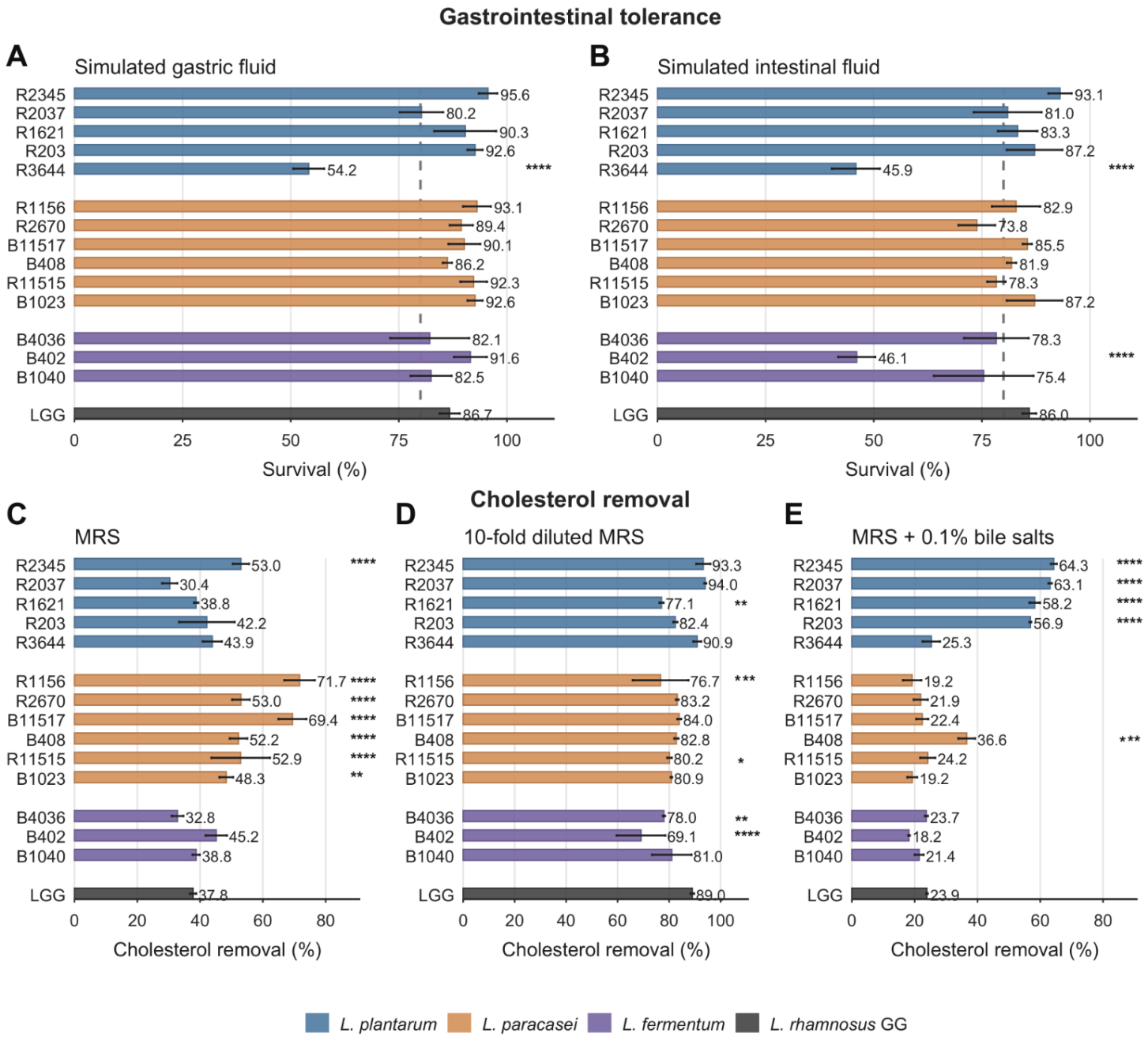


Figure 6

In vitro phenotypes of representative LAB strains. (A, B) Survival after simulated gastric fluid exposure (A) and simulated intestinal fluid exposure (B). (C–E) Cholesterol removal in standard MRS (C), 10-fold diluted MRS (D), and MRS supplemented with 0.1% bile salts (E). Phenotypic data are presented as mean ± SD from three independently prepared culture replicates (n = 3). Asterisks indicate Dunnett-adjusted differences versus LGG: *P < 0.05, **P < 0.01, ***P < 0.001, and ****P < 0.0001

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS3.xlsx
- TableS1.xlsx
- SupplementaryFigureLegends.docx
- TableS2.xlsx
- FigS2.pdf
- FigS3.pdf
- FigS1.pdf