Article

Genomic Discovery of Taxon-Specific Molecular Markers for *Lactobacillaceae* Genera

Sarah Bello and Radhey S. Gupta *

Posted Date: 5 November 2025

doi: 10.20944/preprints202506.2420.v2

Keywords: phylogenomics and comparative genomics; Conserved Signature Indels (CSIs); molecular markers for the *Lactobacillaceae* genera; WGS phylogenetic tree for *Lactobacillaceae*; structural mapping of the CSIs in proteins; TAXIs, taxon-specific indels

 doi:10.20944/preprints202506.2420.v2

Article

Genomic Discovery of Taxon-Specific Molecular Markers for *Lactobacillaceae* Genera

Sarah Bello and Radhey S. Gupta *

Department of Biochemistry and Biomedical Sciences McMaster University, Hamilton, Ontario, CA L8N 3Z5, USA

* Correspondence: gupta@mcmaster.ca

Abstract

Background: Members of the family *Lactobacillaceae*, comprising 36 genera, play vital roles in food fermentation (e.g., wine, yogurt, and cheese production) and contribute significantly to human health through their probiotic properties. Despite their importance, species from different genera are primarily distinguished by phylogenetic clustering and genomic similarity matrices, and no consistent molecular, biochemical, or physiological traits are known that are uniquely found in species from different genera. **Methods**: To address this limitation, we conducted comprehensive phylogenomic and comparative analyses of protein sequences from 410 publicly available *Lactobacillaceae* genomes. **Results:** Based on these analyses, we identified 167 novel conserved signature indels (CSIs) in proteins involved in diverse cellular functions, each specific to a particular genus within the *Lactobacillaceae* family. These taxon-specific CSIs serve as robust molecular markers for genus-level differentiation and have potential applications in functional and diagnostic studies. Using these markers and the AppIndels.com server, we successfully predicted the genus-level affiliation of 111 uncharacterized *Lactobacillus* isolates. Structural analysis of representative CSIs from four genera revealed their consistent location in surface-exposed protein loops, suggesting possible roles in genus-specific protein–protein or protein–ligand interactions. **Conclusions**: The identified CSIs provide novel molecular markers for the robust differentiation of species from different *Lactobacillaceae* genera, offering new tools for exploring the functional traits unique to each genus.

Keywords: phylogenomics and comparative genomics; Conserved Signature Indels (CSIs); molecular markers for the Lactobacillaceae genera; WGS phylogenetic tree for Lactobacillaceae; structural mapping of the CSIs in proteins; TAXIs; taxon-specific indels

1. Introduction

The family *Lactobacillaceae* [1-3] has undergone significant taxonomic revisions in recent years. Notably, *Lactobacillus* species were reclassified into 23 new genera [3] and the former *Leuconostocaceae* family [4,5] was merged into *Lactobacillaceae* [3]. The family *Lactobacillaceae* presently comprises 36 validly published genera [6]. Many *Lactobacillacea*e species from genera such as *Lactobacillus*, *Lacticaseibacillus*, *Lactiplantibacillus*, *Leuconostoc*, *Oenococcus*, and *Weissella*, are widely used in food production and as probiotics due to their health promoting properties [7-10].

Currently, species from different *Lactobacillaceae* genera are distinguished primarily their by clustering in phylogenetic trees based on 16S rRNA genes, core protein sequences, or genomic similarity metrics such as average amino acid identity (AAI), ecological traits, and relative evolutionary distances in the Genome Taxonomy Database (GTDB) [3,11-15]. However, despite extensive research, no consistent biochemical or molecular markers have been identified that reliably differentiate each of the 36 *Lactobacillaceae* genera. The discovery of such markers would not only more reliably differentiate species from each genus but would also provide novel and specific genetic tools for diagnostic and functional studies. Given the ecological diversity and biotechnological relevance of *Lactobacillaceae* species [7-9,11,16-19], new members of this family are being rapidly

 doi:10.20944/preprints202506.2420.v2

discovered [6], with genome sequences increasingly available in public databases such as the NCBI [20]. These resources offer a valuable opportunity for identifying genus-specific molecular markers [20-24].

In our previous work on the former *Leuconostocaceae* family, we identified 46 conserved signature indels (CSIs), or **tax**on-specific **i**ndels (TAXIs), specific to the genera *Convivina*, *Fructobacillus*, *Leuconostoc*, *Oenococcus*, *Periweissella*, and *Weissella* [25]. In our work, the term CSIs refer broadly to all conserved indels specific to different clades, whereas the term TAXIs refer to CSIs which are uniquely associated with a known taxon. These TAXIs, which results from rare genetic changes in a common ancestor of the specific taxa [26-28], serve as molecular synapomorphies, with strong predictive value, and they provide robust evidence of the evolutionary relatedness of species from these lineages [29-31]. Unlike other genus demarcation methods, such as phylogenetic clustering, relative evolutionary distances in the GTDB tree, or genomic similarity metrics like AAI [32], and percentage of conserved proteins (POCP) [21], which can be influenced by various parameters [12,27,33,34], and do not reveal any uniquely shared traits for the species of different genera , TAXIs offer a distinct advantage. These molecular markers are binary presence/absence traits, uniquely found in members of a specific taxon, which make them powerful tools for unambiguous genus-level differentiation [27,31].

Building on our earlier work identifying TAXIs for select *Lactobacillaceae* genera, this study expands the analysis to include additional genera within the family. We analyzed 410 publicly available *Lactobacillaceae* genome sequences (as of July 1, 2024) using phylogenomic and comparative protein sequence analyses. A robust phylogenomic tree was first constructed to delineate evolutionary relationships. We then applied the INDELIBLE **(Indel**-based **I**dentification of **B**acterial **L**ineages and **E**volution) method to identify CSIs specific to individual genera [25,30,35,36]. This analysis led to the discovery of 167 novel TAXIs i.e., genus-specific CSIs in diverse proteins, providing reliable molecular markers for distinguishing between *Lactobacillacea*e genera. To demonstrate their practical utility, we used these TAXIs in conjunction with the AppIndels.com server [37] to predict the taxonomic affiliation of 113 previously uncharacterized *Lactobacillus* isolates. Based on the presence or absence of genus-specific TAXIs, 111 of these strains /isolates were accurately assigned to 11 *Lactobacillaceae* genera.

Structural analysis of representative TAXIs revealed their consistent location in surface-exposed protein loops, suggesting potential functional roles in genus-specific protein–protein or protein–ligand interactions. Further genetic and biochemical studies may uncover novel functional traits, such as metabolic adaptations or ecological specializations [12,19]. Overall, the genus-specific molecular markers identified here enhance taxonomic resolution and provide a foundation for exploring functional diversity across this ecologically and industrially important bacterial family.

2. Materials and Methods

2.1. Construction of Phylogenetic Trees

Genome sequences for type strains and/or reference strains of 410 Lactobacillaceae species, with annotated protein data available in the NCBI database as of July 1, 2024, were downloaded for analysis. To root the phylogenetic tree, genomes of two Bacillus species (Bacillus subtilis and B. cereus) were included as outgroups due to their phylogenetic proximity to Lactobacillaceae [38]. A phylogenomic tree was constructed for our genomes dataset based on concatenated sequences of 87 conserved proteins that comprise the "phyloeco" marker set (single-copy genes universally distributed across members of this phylum) for the phylum Bacillota [39]. The tree was generated using an internally developed pipeline described in our earlier work [31,40]. Briefly, homologs of the 87 phyloeco set of proteins were identified using the CD-HIT v4.6 program and profile Hidden Markov Models of these proteins [41], at default settings. Proteins present in at least 80% of the genomes and sharing a minimum of 50% sequence identity and length were retained. Multiple sequence alignments of these protein families were generated using the Clustal Omega v1.2.2 at

 doi:10.20944/preprints202506.2420.v2

default settings [42], and poorly aligned regions were removed using trimAl v1.4 [43]. The final concatenated sequence alignment consisted of 26527 aligned positions. A maximum-likelihood phylogenetic tree based on this alignment was constructed using the Whelan and Goldman model of protein evolution [44] and formatted using MEGA X [45].

2.2. Identification of Conserved Signature Indels (CSIs)

Identification of CSIs for different *Lactobacillaceae* genera was performed using methods described in our earlier studies [25,27,35]. Briefly, local BLASTp searches were conducted on protein sequences from genomes of representative *Lactobacillus* species across different clades of interest. Based on these searches, sequences of high-scoring homologs (E value <1e-^{20}) were retrieved for 4 to 10 within the target group and 10-15 outgroup species. Multiple sequence alignments were generated using Clustal X 2.1 [46] and visually inspected for the presence of insertions or deletions (indels) of fixed lengths located within conserved regions. Indels that were retained for further analysis met the following criteria [25,27,35]: (i) flanked on both sides by at least 5–6 conserved amino acid residues within a 40–50 residue window, and (ii) primarily found in species from a specific *Lactobacillaceae* genus but absent in other genera and outgroup species. Indels not located in conserved regions were excluded [25,27,35].

To confirm their taxon specificity, candidate CSIs, along with 30–50 flanking amino acids, were used in broader BLASTp searches against the NCBI non-redundant (nr) database. The top 300–500 hits were examined to assess the taxonomic distribution of each indel. Indels found exclusively in species from a single *Lactobacillaceae* genus were designated as CSIs and formatted using the SIGCREATE and SIGSTYLE tools [27,35], available via the Gleans.net server (http://gleans.net/). Due to space limitations, representative sequence data are shown in the main figures for selected species. Unless otherwise noted, all described CSIs are genus-specific and absent from other bacterial homologs among the top BLASTp hits. Additional sequence details for each CSI are provided in the Supplemental Figures.

2.3. Taxonomic Predictions of Lactobacillus Strains/Isolates Using AppIndels.com Sever

Protein sequences for 113 uncharacterized *Lactobacillus* isolates were retrieved from the NCBI genome database in .faa format. These sequences were analyzed using the AppIndels.com server [37] which performs local BLASTp searches to detect TAXIs that are specific for different genera. If the number of TAXIs identified in a genome exceeds the predefined threshold for a given taxon, the server assigns the genome to that taxon. A detailed description of the server's methodology is available in previous work [37].

2.4.Determination of Protein Structures Using AlphaFold Model Generation to Map the Locations of CSIs

To investigate the structural context of CSIs, we analyzed four proteins containing genus-specific CSIs from *Apilactobacillus, Lacticaseibacillus, Lactiplantibacillus*, and *Lactobacillus*. FASTA sequences of both CSI-containing proteins and their homologs lacking the CSIs were retrieved from the NCBI protein database and submitted to the AlphaFold 3 server for structure prediction using default parameters [47]. The top-ranked predicted models were visualized and analyzed using PyMOL v2.5.5 [48]. Structural confidence was assessed using predicted Local Distance Difference Test (pLDDT) and predicted Template Modeling (pTM) scores [49,50]. Only models with high-confidence predictions (pLDDT > 50 and pTM > 0.8) were included in further analyses [47,51]. Final models were superimposed in PyMOL to localize the CSIs within the protein structures [48]. Structural similarity be-tween CSI-containing and CSI-lacking homologs was evaluated using root mean square deviation (RMSD) values.

3. Results

3.1. Phylogenomic Tree for the Lactobacillaceae Species

 doi:10.20944/preprints202506.2420.v2

To determine the evolutionary relationships and genus-level affiliations of all 410 *Lactobacillaceae* species with available genomes as of July 1, 2024, we constructed a maximum likelihood phylogenomic tree using concatenated sequences of 87 conserved proteins. The resulting tree is shown in a compressed form in Fig. 1, where species clades from different genera are coalesced for clarity. However, an expanded, uncompressed version of this tree is provided in Fig. S1. In this tree (Fig.1 and Fig. S1), which we will be referred to as the phyloeco tree, nearly all nodes exhibit 100% statistical support, indicating strong confidence in the inferred relationships. All examined *Lactobacillaceae* species clustered within clades corresponding to their respective genera, with branching patterns consistent with previous studies [3,12]. Species from several *Lactobacillaceae* genera (viz. *Acetilactobacillus*, *Amylolactobacillus*, *Philodulcilactobacillus*, *Holzapfeliella*, *Nicoliella*, *Paralactobacillus*), which contain only a single species also showed distinct branching, supporting their classification as separate genera [3].

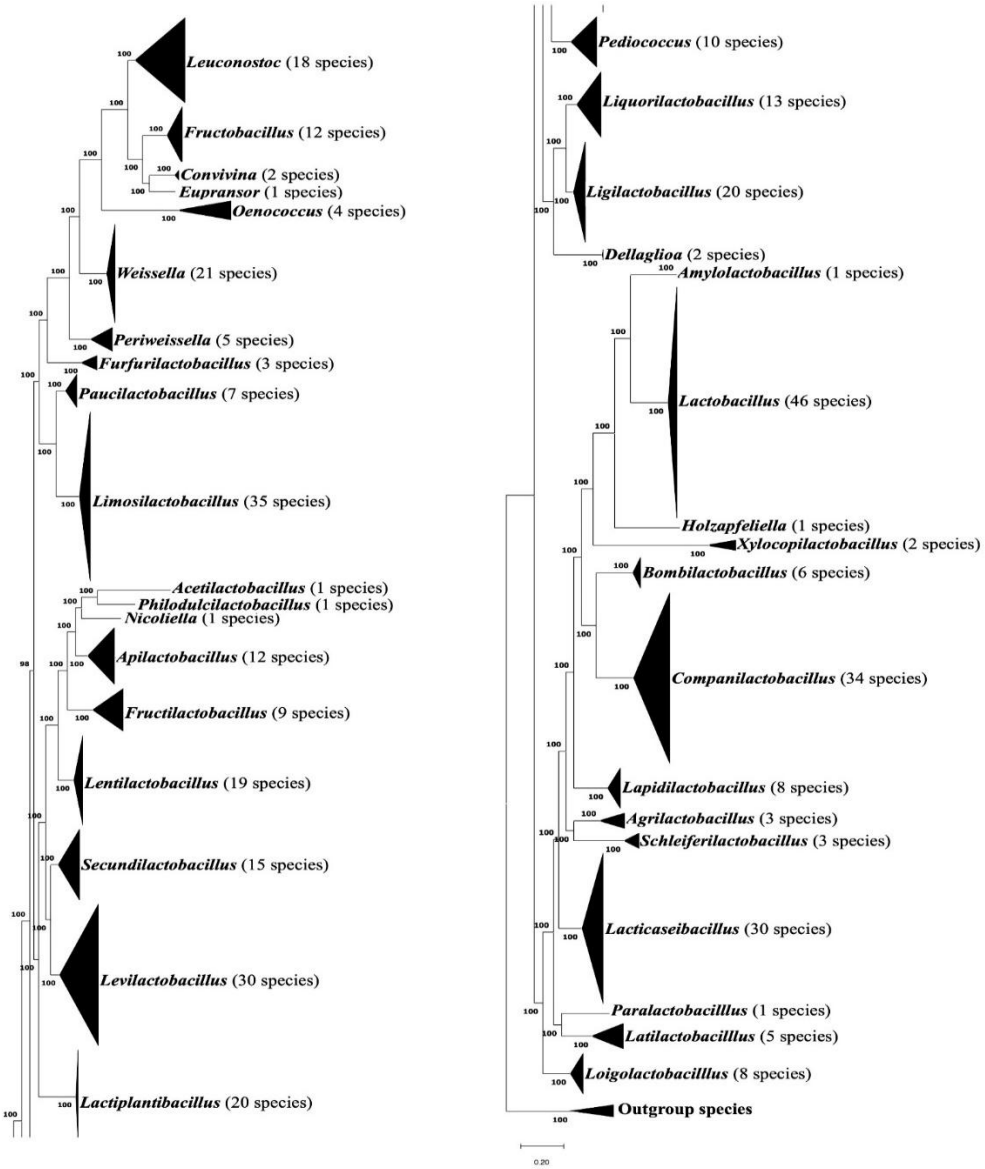

Figure 1. A bootstrapped maximum-likelihood tree for 410 genome-sequenced *Lactobacillaceae* species based on concatenated sequences of 87 conserved proteins. The statistical support values for different branches are indicated on the nodes. This tree was rooted by using *Bacillus* species as an outgroup (see Methods). Different main species clades in this tree are identified by the names of the genera and are compressed. An uncompressed version of this tree is presented in Fig. S1.

3.2. Conserved Signature Indels Specific for Different Lactobacillaceae Genera

 doi:10.20944/preprints202506.2420.v2

The phylogenetic tree shown in Figure 1 (and expanded in Figure S1) provides strong support for the monophyly of the various *Lactobacillaceae* genera. This tree forms the foundation for the central focus of this study, i.e. the identification of TAXIs unique to individual genera. Previous research across diverse prokaryotic taxa has demonstrated the value of TAXIs as reliable molecular markers for evolutionary and taxonomic studies [25,27,31,35,40,52-54]. Building on this foundation, we conducted detailed analyses of protein sequences from *Lactobacillaceae* species using the INDELIBLE approach. These analyses have identified167 novel CSIs in diverse proteins. Each of these CSI is uniquely present in species from a specific *Lactobacillaceae* genus. The results of these analyses are summarized and discussed below.

3.3. Molecular Markers Specific for the Genera Lactobacillus, Lacticaseibacillus Lactiplantibacillus and Apilactobacillus

The genus *Lactobacillus* is the type genus of the family *Lactobacillaceae* and remains its most populous and extensively studied member [3,55]. Prior to its reclassification in 2020, the genus included over 260 species, many of which exhibited polyphyletic branching alongside species from the former *Leuconostocaceae* family, and displayed substantial phenotypic and ecological diversity [12,13,16,56,57]. However, following the taxonomic revision by Zhang et al. [3], species from the genus *Lactobacillus* were redistributed into 23 distinct genera, resulting in a monophyletic grouping of different proposed genera in phylogenetic analyses. The composition and branching of species within the genus *Lactobacillus*, as observed in our phylogenetic tree, are shown in Figure 2A.

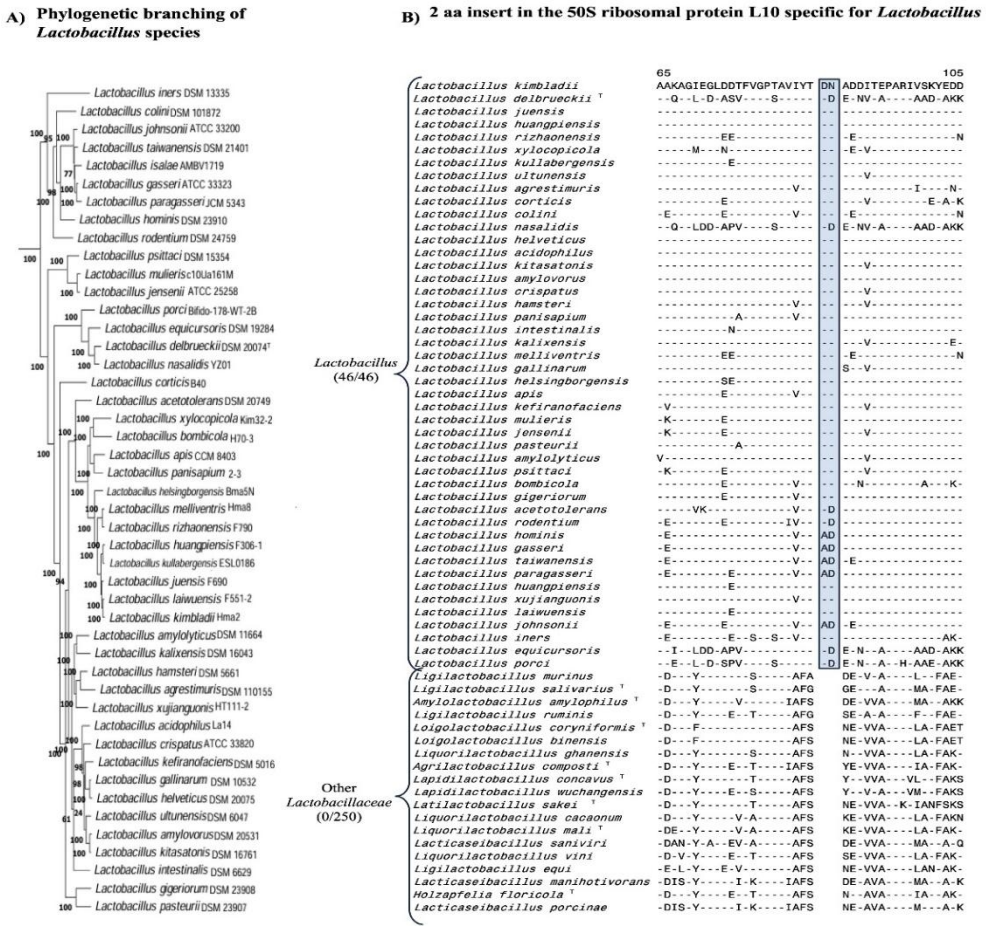

Figure 2. (A) Branching pattern of species from the genus *Lactobacillus* in our phylogenomic tree. (B) Partial sequence alignment of the 50S ribosomal protein L10 showing a two amino acid insertion (highlighted) uniquely shared by species/strains from the genus *Lactobacillus*. Dashes (–) indicate identity with the amino acids in the top reference sequence, while gaps represent missing residues at those positions. Accession numbers for each

 doi:10.20944/preprints202506.2420.v2

sequence are listed in the second column, and the position of the sequence fragment within the protein is shown above the alignment. Detailed sequence data for this CSI and 15 additional *Lactobacillus*-specific CSIs are provided in Figures S2–S17, with a summary of their characteristics in Table 1.

Despite its division multiple genera, *Lactobacillus* still comprises over 46 named species [6], which exhibit considerable genetic diversity (Fig. 2A). Several species from this genus are widely used as probiotics [18,58,59], while others such as *Lactobacillus delbrueckii* subsp. *bulgaricus*, play key roles in dairy fermentation including yogurt production [60,61]. However, despite the industrial and scientific significance, no molecular characteristics have previously been identified that are uniquely specific to *Lactobacillus* species.

Our analyses identified 16 novel CSIs in diverse proteins, most of which are uniquely shared by all or most *Lactobacillus* species. One example is shown in Fig. 2B, where a two aa insertion within a conserved region of the 50S ribosomal protein subunit L10 is found exclusively in all 46 genome-sequenced *Lactobacillus* species. This protein is part of the L7/L12 stalk of the 50S ribosomal subunit and plays a key role in protein synthesis by recruiting translation factors and stimulating GTP hydrolysis [62]. In this and other sequence alignments, dashes (–) indicate identity with the amino acid shown on the top line. This CSI is absent in all other *Lactobacillaceae* species and in other examined bacteria. Due to space constraints, Fig. 2B displays sequence data for a subset of *Lactobacillus* and other *Lactobacillaceae* species. Comprehensive sequence data for this and the remaining 15 CSIs, each found in different proteins, are provided in Figs. S2–S17, with key features summarized in Table 1. Given their specificity, these TAXIs likely originated in a common ancestor of the *Lactobacillus* genus and serve as reliable molecular markers for distinguishing its species from those of other *Lactobacillaceae* genera.

The genus *Lacticaseibacillus* comprises species that have been extensively studied for their probiotic properties, their involvement in dental caries, and their growing association with bacteremia [63,64]. In our phylogenetic analysis (Fig. S1), members of this genus form a well-supported, deeply branching monophyletic clade. The species composition and branching pattern within this clade are illustrated in Fig. 3A. Using the INDELIBLE approach, we identified nine novel conserved CSIs in various proteins that are uniquely present in *Lacticaseibacillus* species. One representative example is shown in Fig. 3B, where a one amino acid insertion in a conserved region of the manganese-dependent inorganic pyrophosphatase protein is found in all 30 genome-sequenced *Lacticaseibacillus* species but absent in other *Lactobacillaceae* genera. Detailed sequence information for this CSI, along with the eight additional TAXIs identified for this genus, is provided in Figs. S18–S26. Key features of these CSIs are summarized in Table 1. Together, these molecular markers offer reliable tools for distinguishing *Lacticaseibacillus* species from other members of the *Lactobacillaceae* family.

The genus *Apilactobacillus* comprises 12 validly published species, which form a distinct clade in our phylogenomic tree (Fig. 1). The species composition and branching pattern within this clade are shown in Fig. 4A. Members of this genus are predominantly associated with fructose-rich environments, such as the guts of bees and flowers, highlighting their ecological link to insects [3,12,65,66]. Our analysis identified four CSIs that are uniquely present in *Apilactobacillus* species. One representative example is shown in Fig. 4B, where a two–amino acid insertion in the cyclopropane-fatty-acyl-phospholipid synthase family protein is uniquely shared by all 12 *Apilactobacillus* species but absent in other *Lactobacillaceae* genera. Detailed sequence information for this CSI along with the other three *Apilactobacillus*-specific CSIs is provided in Figs. S27–S30. Key characteristics of these CSIs/TAXIs are summarized in Table 1. These TAXIS offer reliable tools for distinguishing *Apilactobacillus* species from other *Lactobacillaceae* genera.

Species of the genus *Lactiplantibacillus* inhabit a wide range of environments, including fermented foods (e.g., sauerkraut, kimchi), plant material, and the human gastrointestinal tract [3]. Among them, *Lactiplantibacillus plantarum* has been extensively studied for its probiotic benefits, including its ability to ferment plant-derived and phenolic compounds, its antioxidant properties,

 doi:10.20944/preprints202506.2420.v2

and its antimicrobial activity through bacteriocin production [67,68]. The genus *Lactiplantibacillus* currently includes 20 validly published species, which form a well-supported clade in the phylogenomic tree constructed in this study. The branching pattern of species from this genus is shown in Fig. 4C.

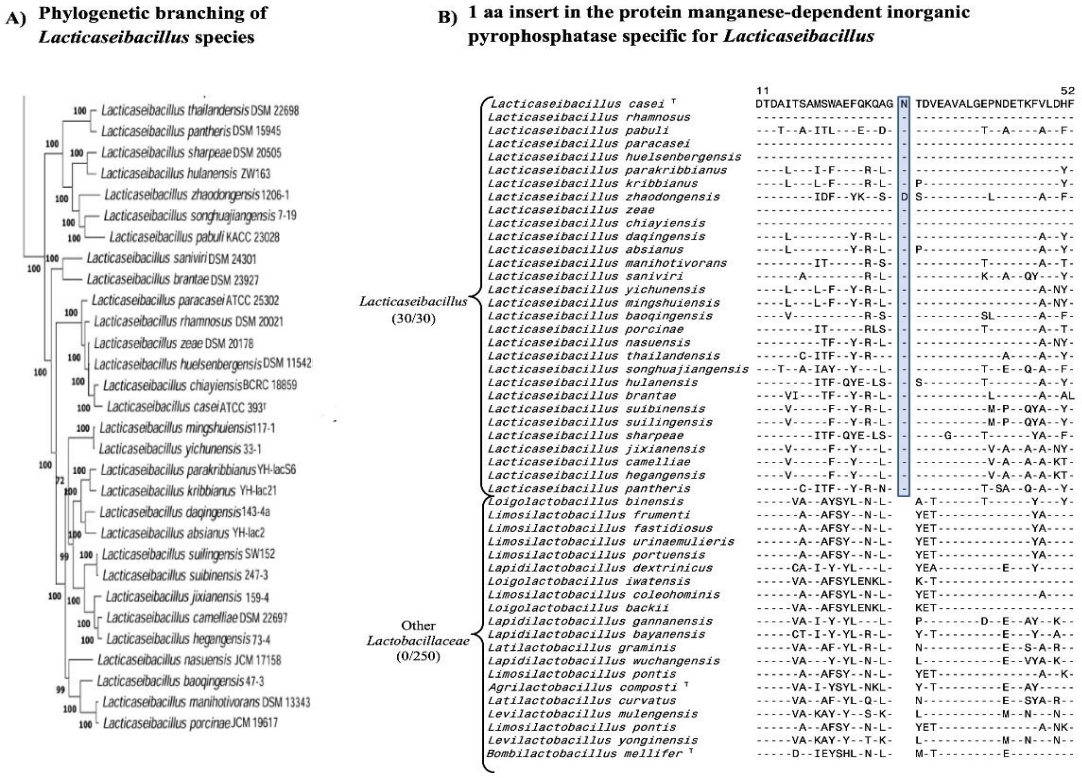

Figure 3. (A) Branching pattern of *Lacticaseibacillus* species in our phylogenomic tree. (B) Excerpt from the sequence alignment of the manganese-dependent inorganic pyrophosphatase protein showing a one amino acid insertion uniquely shared by species/strains of the genus *Lacticaseibacillus*. Detailed sequence data for this CSI, along with eight additional *Lacticaseibacillus*-specific CSIs, are provided in Figures S18–S26, with a summary of their characteristics in Table 1.

 doi:10.20944/preprints202506.2420.v2

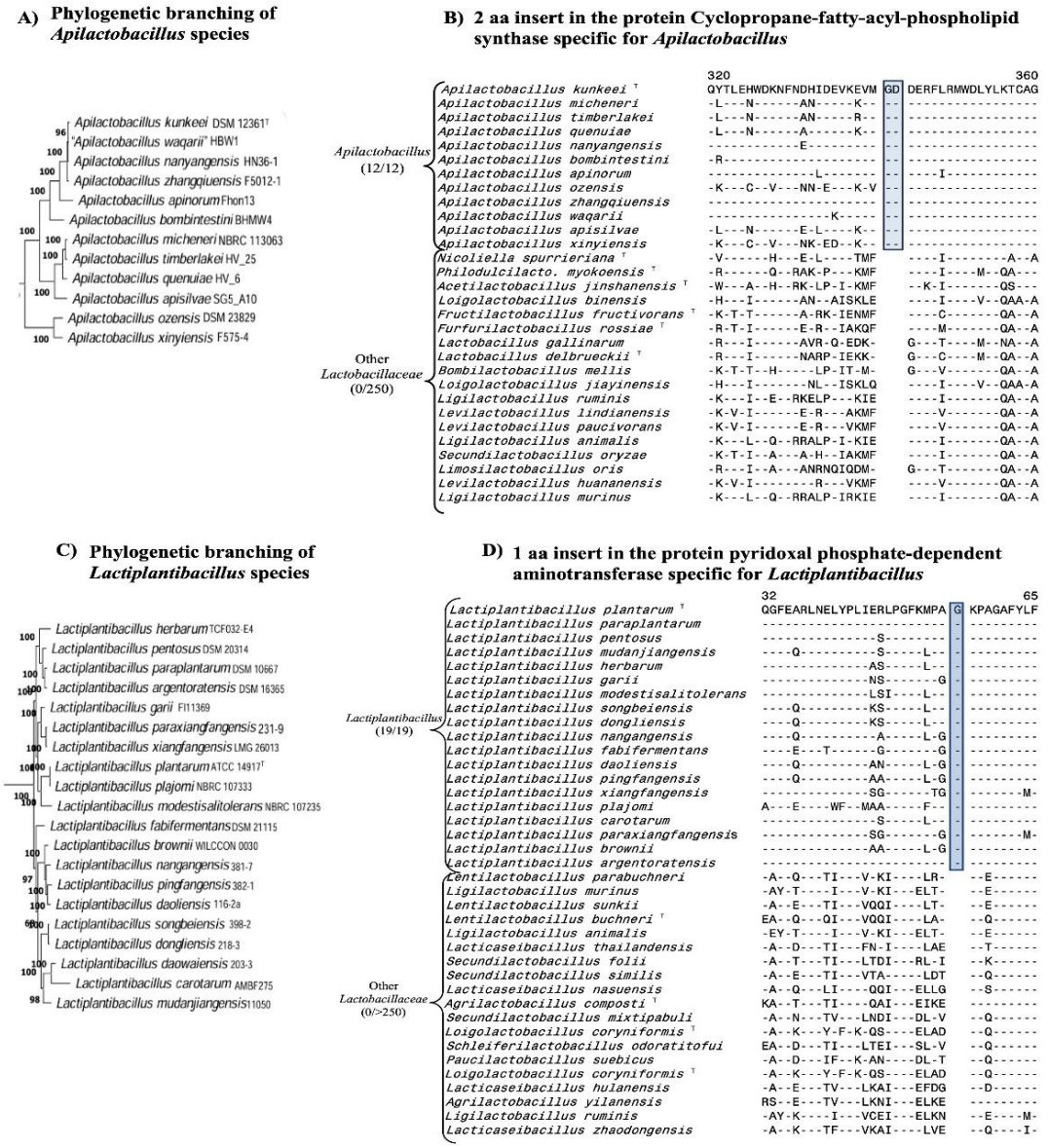

Figure 4. Branching patterns of *Apilactobacillus* (A) and *Lactiplantibacillus* (C) species in our phylogenomic tree. Examples of molecular signatures specific to the genera *Apilactobacillus* (B) and *Lactiplantibacillus* (D). (A) Partial sequence alignment of the cyclopropane-fatty-acyl-phospholipid synthase family protein showing a two amino acid insertion uniquely shared by species/strains of *Apilactobacillus*. Detailed sequence data for this CSI and four additional *Apilactobacillus*-specific CSIs are shown in Figures S27–S30, with a summary in Table 1. (B) Excerpt from the sequence alignment of pyridoxal phosphate-dependent aminotransferase showing a one amino acid insertion uniquely shared by species of *Lactiplantibacillus*. Detailed sequence data for this CSI and seven additional CSIs for this genus are presented in Figures S31–S38, with their characteristics summarized in Table 1.

Our comparative genomic analysis identified eight CSIs in various proteins that are uniquely present in species of the genus *Lactiplantibacillus*. One representative CSI is shown in Fig. 4D, where a single amino acid insertion in a highly conserved region of the pyridoxal phosphate-dependent aminotransferase protein is uniquely shared by all *Lactiplantibacillus* species and absent in all other *Lactobacillaceae* genera. Detailed sequence information for this CSI, along with the other seven *Lactiplantibacillus*-specific CSIs, is provided in Figs. S31–S38. Key characteristics of these markers are summarized in Table 1. These TAXIS provide reliable molecular tools for distinguishing *Lactiplantibacillus* species from all other *Lactobacillaceae* genera.

 doi:10.20944/preprints202506.2420.v2

Table 1. Summary of CSIs specific for the genus *Lactobacillus*, *Lacticaseibacillus*, *Apilactobacillus* and *Lactiplantibacillus*.

Protein Name	Accession No.	Indel Size	Indel Position	Figure No.	Specificity
50S ribosomal protein L10	WP_046332409	2 aa Ins	57-112	**Fig. 2** Fig. S2	*Lactobacillus*
excinuclease ABC subunit UvrC	WP_0036197795	-6 aa Ins	480-531	Fig. S3	
Anaerobic ribonucleoside-triphosphate reductase§	WP_011161356	2 aa Ins	517-562	Fig. S4	
DNA-binding protein WhiA	WP_004893933	1 aa Ins	140-194	Fig. S5	
Translation initiation factor IF-2	WP_011544002	3 aa Ins	285-336	Fig. S6	
50S ribosomal protein L4	WP_046332456	2 aa Del	120-280	Fig. S7	
TIGR01457 family HAD-type hydrolase	WP_046331702	1 aa Ins	98-130	Fig. S8	
C69 family dipeptidase *§	WP_003647856	1 aa Del	345-389	Fig. S9	
YfbR-like 5'-deoxynucleotidase§	WP_057718391	1 aa Ins	23-79	Fig. S10	
class I SAM-dependent methyltransferase	WP_003619061	1 aa Del	269-326	Fig. S11	
Phosphate acyltransferase PlsX§	WP_011162257	1 aa Del	176-227	Fig. S12	
DNA helicase PcrA*§	WP_011162397	2 aa Ins	248-301	Fig. S13	
NADP-dependent phosphogluconate dehydrogenase *§	WP_011162624	1 aa Del	5-57	Fig. S14	
calcium-translocating P-type ATPase§	WP_044025971	1 aa Del	814-864	Fig. S15	
ATP-binding protein*§	WP_046332316	1 aa Ins	347-399	Fig. S16	
16S rRNA (cytosine(1402)-N(4))-methyltransferase RsmH*§	WP_044496740	1 aa Ins	76-113	Fig. S17	
manganese-dependent inorganic pyrophosphatase*	WP_003579130	1 aa Ins	9-59	**Fig. 3** Fig. S18	*Lacticaseibacillus*
hemolysin family protein	WP_138426554	1 aa Ins	345-382	Fig. S19	
1-acyl-sn-glycerol-3-phosphate acyltransferase	WP_049169464	1 aa Del	142-191	Fig. S20	
DUF1002 domain-containing protein*§	WP_049172803	1 aa Del	85-129	Fig. S21	
DeoR/GlpR family DNA-binding transcription regulator*	WP_191995078	1 aa Del	85-128	Fig. S22	
DNA polymerase IV	WP_138131441	1 aa Del	110-155	Fig. S23	
DNA polymerase IV*	WP_138131441	1 aa Del	227-263	Fig. S24	
YfcE family phosphodiesterase*	WP_129319710	1 aa Del	1-36	Fig. S25	
methionine adenosyltransferase	WP_138426285	1 aa Del	58-102	Fig. S26	
cyclopropane-fatty-acyl-phospholipid synthase family protein	WP_138741898	2 aa Ins	315-362	**Fig. 4(B)** Fig. S27	*Apilactobacillus*
DEAD/DEAH box helicase	WP_053791914	1 aa Ins	168-209	Fig. S28	
Phosphate acetyltransferase*	WP_053791569	1 aa Del	200-239	Fig. S29	
glucose-6-phosphate dehydrogenase	WP_053796109	1 aa Ins	12-48	Fig. S30	
pyridoxal phosphate-dependent aminotransferase	WP_208215537	1 aa Ins	30-65	**Fig. 4(D)** Fig. S31	*Lactiplantibacillus*
ABC transporter ATPase §	KLD61660	1 aa Del	44-98	Fig. S32	
acetyl-CoA carboxylase §	KLD60369	1 aa Ins	32-83	Fig. S33	
50S ribosomal protein L15§	WP_021337917	1 aa Del	83-126	Fig. S34	
C69 family dipeptidase	WP_134144186	2 aa Del	289-325	Fig. S35	
GRP family sugar transporter §	WP_222843328	1 aa Del	83-128	Fig. S36	
glycoside hydrolase family 13 protein*	WP_064619115	1 aa Del	377-430	Fig. S37	

 doi:10.20944/preprints202506.2420.v2

undecaprenyl-phosphate alpha-N-acetylglucosaminyl 1-phosphate transferase §	OAX76783	1 aa Del	158-208	Fig. S38

*. Isolated exceptions present in ingroup and/or outgroup species. §. Protein homolog is missing from ingroup and/or outgroup species.

Figures 2–4 present selected examples of TAXIs identified for some *Lactobacillaceae* genera. However, in addition to the results shown for *Lactobacillus*, *Lacticaseibacillus*, *Apilactobacillus*, and *Lactiplantibacillus*, our comprehensive protein sequence analyses across other *Lactobacillaceae* genera have revealed an additional 130 novel CSIs. Most of these CSIs are uniquely shared by species within a single genus, making them reliable molecular signatures for genus-level identification.

The numbers of CSIs that we identified for the other *Lactobacillaceae* genera in this study are as follows: *Agrilactobacillus* (4), *Amylolactobacillus* (4), *Bombilactobacillus* (7), *Companilactobacillus* (10), *Dellaglioa* (6), *Fructilactobacillus* (8), *Furfurilactobacillus* (19), *Lapidilactobacillus* (4), *Latilactobacillus* (8), *Lentilactobacillus* (3), *Levilactobacillus* (4), *Ligilactobacillus-Liquorilactobacillus* cluster (7) *Limosilactobacillus* (8), *Loigolactobacillus* (6), *Paucilactobacillus* (3), *Pediococcus* (10), *Schleiferilactobacillus* (15), *Secundilactobacillus* (4), and *Xylocopilactobacillus* (4).

It should be noted that although our analysis identified multiple CSIs for all *Lactobacillaceae* genera containing two or more species, no genus-specific CSIs were found for *Ligilactobacillus* and *Liquorilactobacillus*. Previous studies indicate that species from these two genera, comprising primarily of free-living bacteria and vertebrate-associated species, respectively, are closely related [3,69]. Supporting this inference, our analysis has also identified eight CSIs that are uniquely shared between species of both genera, suggesting they shared a common ancestor distinct from other *Lactobacillaceae*.

Detailed sequence information for CSIs specific for the other *Lactobacillaceae* genera is provided in Figs. S39–S168, and key characteristics are summarized in Tables S1–S4. In addition to these findings, our previous work identified multiple CSIs specific for several other *Lactobacillaceae* genera[25], which were formerly classified under the family *Leuconostocaceae* [5]. The numbers of CSIs identified for these genera are as follows: *Fructobacillus* (5), *Leuconostoc* (5), *Oenococcus* (13), *Periweissella* (5), and *Weissella* (6) [25]. A summary of the species compositions of the various *Lactobacillaceae* genera, including those formerly classified under the family *Leuconostocaceae*, and the number of genus-specific CSIs (TAXIs) identified for each is presented in Fig. 5. Based on these findings, all *Lactobacillaceae* genera containing two or more species, except *Ligilactobacillus* and *Liquorilactobacillus*, can now be reliably distinguished from one another using multiple, genus-specific TAXIs.

 doi:10.20944/preprints202506.2420.v2

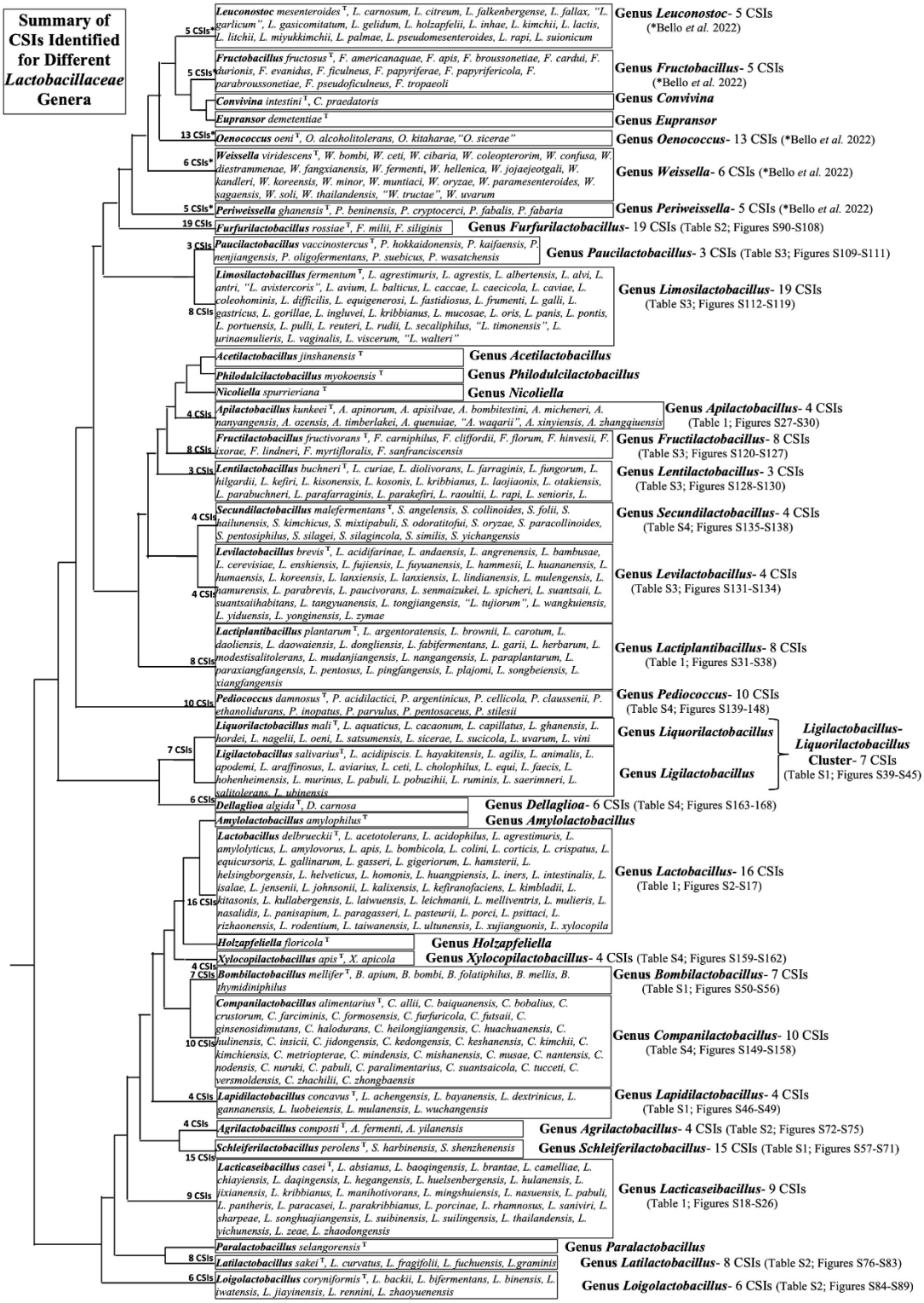

Figure 5. Summary diagram showing the species composition of various *Lactobacillaceae* genera and the number of taxon-specific CSIs identified for each. Asterisks (*) indicate CSIs previously reported in Bello et al. (2022)[25].

3.4. Predictive Ability of Previously Identified CSIs to be Found in Newly Described Species

Previous studies on CSIs specific to various taxa and genera have demonstrated that they exhibit strong predictive value i.e., once identified in known members of a group, these markers are often consistently found in newly sequenced or discovered members of the same lineage [31,37,70-72]. This predictive capability is further illustrated in Fig. 6, which presents updated sequence data for two CSIs specific to *Leuconostocaceae* genera identified in our previous work [25].

 doi:10.20944/preprints202506.2420.v2

Fig. 6A shows an eight aa insertion in the protein phospho-N-acetylmuramoyl-pentapeptide-transferases, originally identified as specific to the genus *Weissella* [25], whose members are known for their probiotic and biotechnological potential [73]. At the time of its discovery, sequence data were available for 14 *Weissella* species. Since then, genome sequences for four additional species have become available, and all of them contain this CSI, highlighting its stability and predictive value.

Similarly, Fig. 6B shows a CSI specific to the genus *Fructobacillus*, a group of fructose-fermenting microorganisms [74]. When this CSI was first reported in 2022, it was found in five species. Since then, eight new *Fructobacillus* species have been described [6], and this CSI is present in all of them. These results provide compelling evidence highlighting the long-term stability and genus-specific conservation of CSIs, reinforcing their utility as reliable molecular markers for taxonomic classification and evolutionary studies.

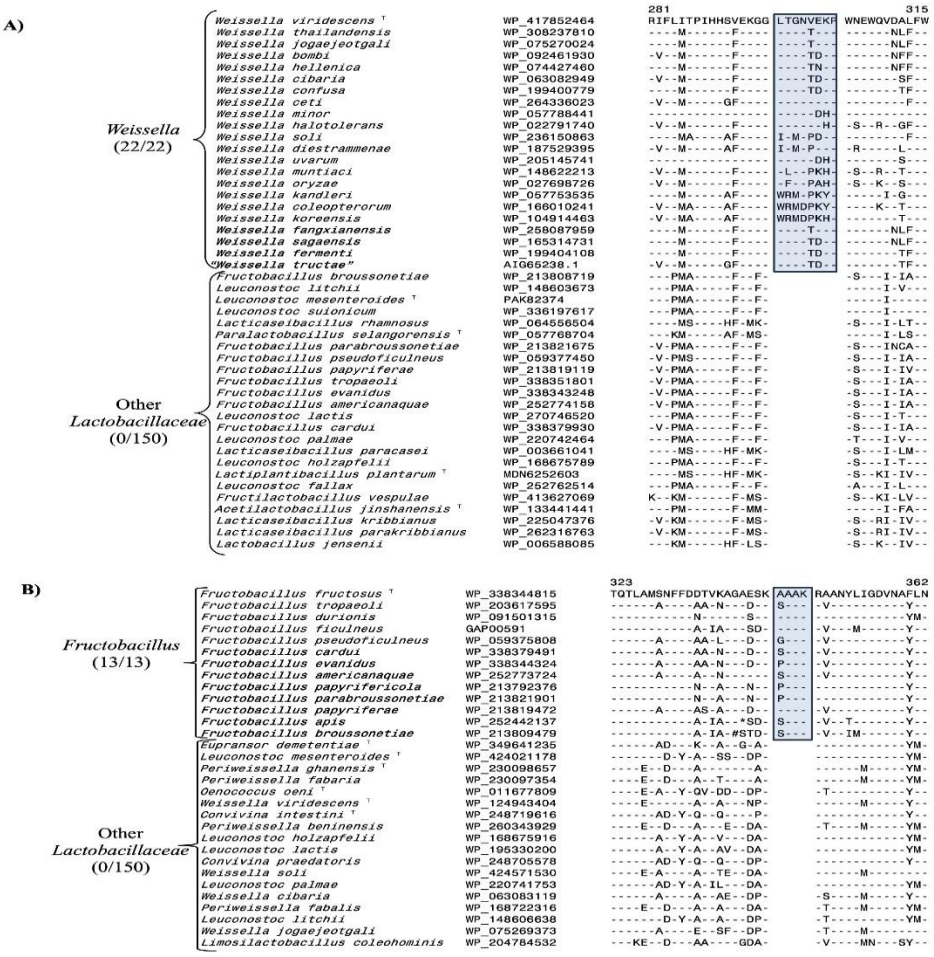

Figure 6. Updated sequence alignments of molecular signatures specific to the genera *Weissella* and *Fructobacillus*, originally described in Bello et al. [25]. This figure has been adapted to include newly sequenced species from both genera. (A) Excerpt from the alignment of the phospho-N-acetylmuramoylpentapeptide-transferase protein showing an eight–amino acid insertion in a conserved region uniquely shared by all *Weissella* species. (B) Excerpt from the alignment of the Asp-tRNA(Asn)/Glu-tRNA(Gln) amidotransferase subunit (GatB) showing a four–amino acid insertion specific to all *Fructobacillus* species. * and # denote additional aa residues are in these positions.

3.5. Application of the Identified CSIs for Taxonomic Prediction of Uncharacterized Lactobacillus isolates

Based on the predictive capability of the TAXIs, we have recently developed a web-based tool, AppIndels.com, which uses the presence of known TAXIs in genome sequences to predict taxonomic affiliations [37]. To evaluate the utility of the TAXIs identified in this study for *Lactobacillaceae* genera, we added the corresponding CSI sequence data to the AppIndels.com server and used it to analyze

 doi:10.20944/preprints202506.2420.v2

113 uncharacterized *Lactobacillus* isolates with available genome sequences in the NCBI database. In Figure S169, results from the server are shown for two representative *Lactobacillus* isolates. For *Lactobacillus* sp. CBA3605, the server predicted affiliation with the genus *Lactiplantibacillus*, identifying eight CSIs specific to this clade (Fig. S169A). In contrast, *Lactobacillus* sp. UW*DM*LACCAS1_1 was predicted to belong to the genus *Lacticaseibacillus*, with nine genus-specific CSIs detected in its genome (Fig. S169B). In addition to reporting the number of CSIs matching the predicted genus, the server provides access to the corresponding sequence data[37], which can be viewed by clicking the arrow next to the CSI count.

The taxonomic predictions made by the AppIndels.com for the genome sequences of all examined *Lactobacillus* isolates are summarized in Table S5. This table includes the accession numbers of the genomes, the predicted taxonomic affiliations from AppIndels.com, and the number of TAXIs identified for each genome.

As seen from the results presented in Table S5, the server successfully predicted genus-level affiliations for 111 out of 113 isolates based on the presence of multiple genus-specific TAXIs. These isolates were assigned to the following 10 genera: *Agrilactobacillus* (1), *Bombilactobacillus* (7), *Fructilactobacillus* (1), *Lacticaseibacillus* (8), *Lactiplantibacillus* (2), *Lactobacillus* (81), *Lentilactobacillus* (1), *Levilactobacillus* (2), *Limosilactobacillus* (7), and *Ligilactobacillus-Liquorilactobacillus* cluster (1). For two genomes (GCA_014796685.1 and GCA_019303535.1) no taxonomic predictions were made by the server. One of these genomes (accession number GCA_014796685.1) is indicated as contaminated in its NCBI record.

To validate the accuracy of these predictions, we constructed a phylogenetic tree that includes the 111 uncharacterized isolates along with representative species from various *Lactobacillaceae* genera (Fig. 7). As seen from the tree in Fig. 7, there was 100% concordance between the genus assignments predicted by the AppIndels.com server and the phylogenetic placements of the isolates. It should be noted that the high accuracy of the AppIndels.com server in predicting taxonomic affiliation results from its requirement that it predicts a positive taxonomic affiliation only when multiple CSIs specific for a particular genus are present in the analyzed genome [75]. Since each TAXI represents a rare genetic change [27,28,35], the likelihood of multiple CSIs from the same genus appearing in a genome by chance is extremely low. This feature of the AppIndels.com server ensures that the taxonomic predictions made by it are accurate. These results demonstrate that the TAXIs identified in this study provide a robust and practical tool for determining the taxonomic affiliation of novel or uncharacterized isolates from this family.

 doi:10.20944/preprints202506.2420.v2

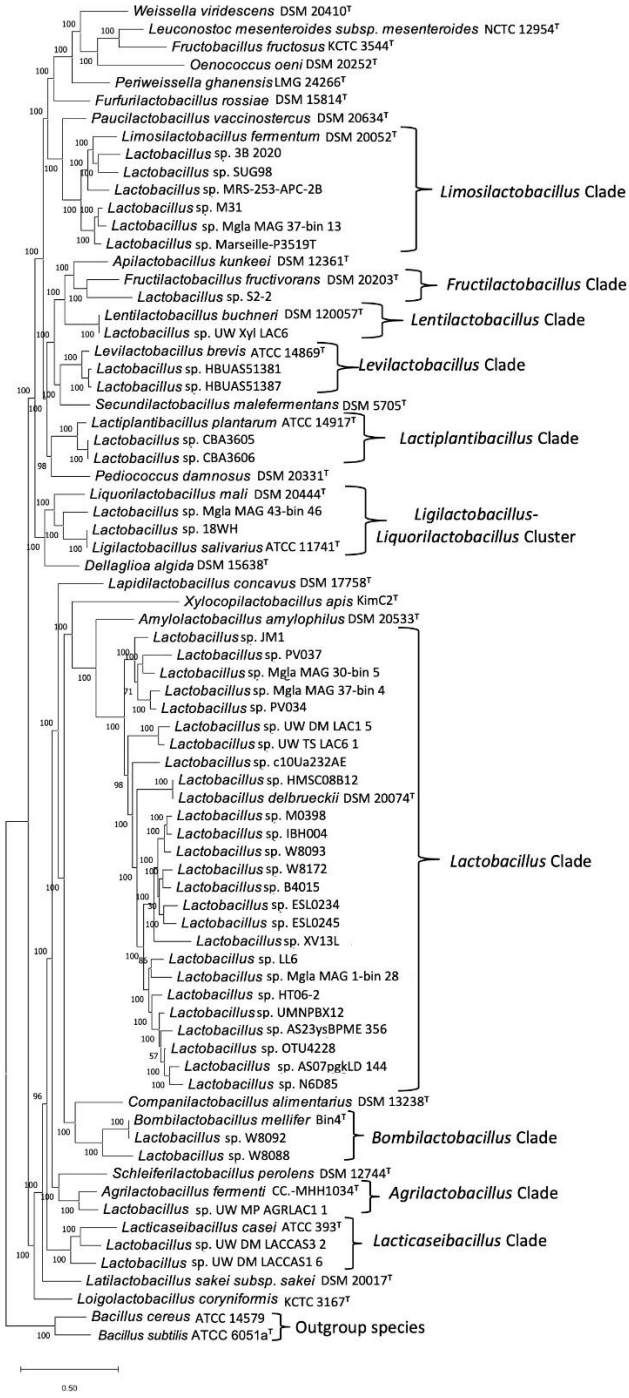

Figure 7. A bootstrapped maximum-likelihood tree showing the branching of the type species of various *Lactobacillaceae* genera along with uncharacterized *Lactobacillus* isolates which were predicted to correspond to specific genera by the AppIndels.com server. Due to space constraints, some closely related strains are not shown. Clades for different *Lactobacillaceae* genera and the associated uncharacterized isolates are labeled in the tree.

3.6. Taxon-specific CSIs are Localized in Surface Exposed Loops of Proteins

Previous studies on CSIs specific to various prokaryotic taxa have shown that these genetic changes are frequently located in surface-exposed loop regions of proteins, which are flexible, unstructured areas on solvent-accessible surfaces that often mediate novel protein–protein or

 doi:10.20944/preprints202506.2420.v2

protein–ligand interactions [76-81]. Considering these findings, we investigated the structural localization of selected CSIs specific to *Lactobacillaceae* genera that shown in Figs. 2–4. Results of these analyses are presented in Fig. 8.

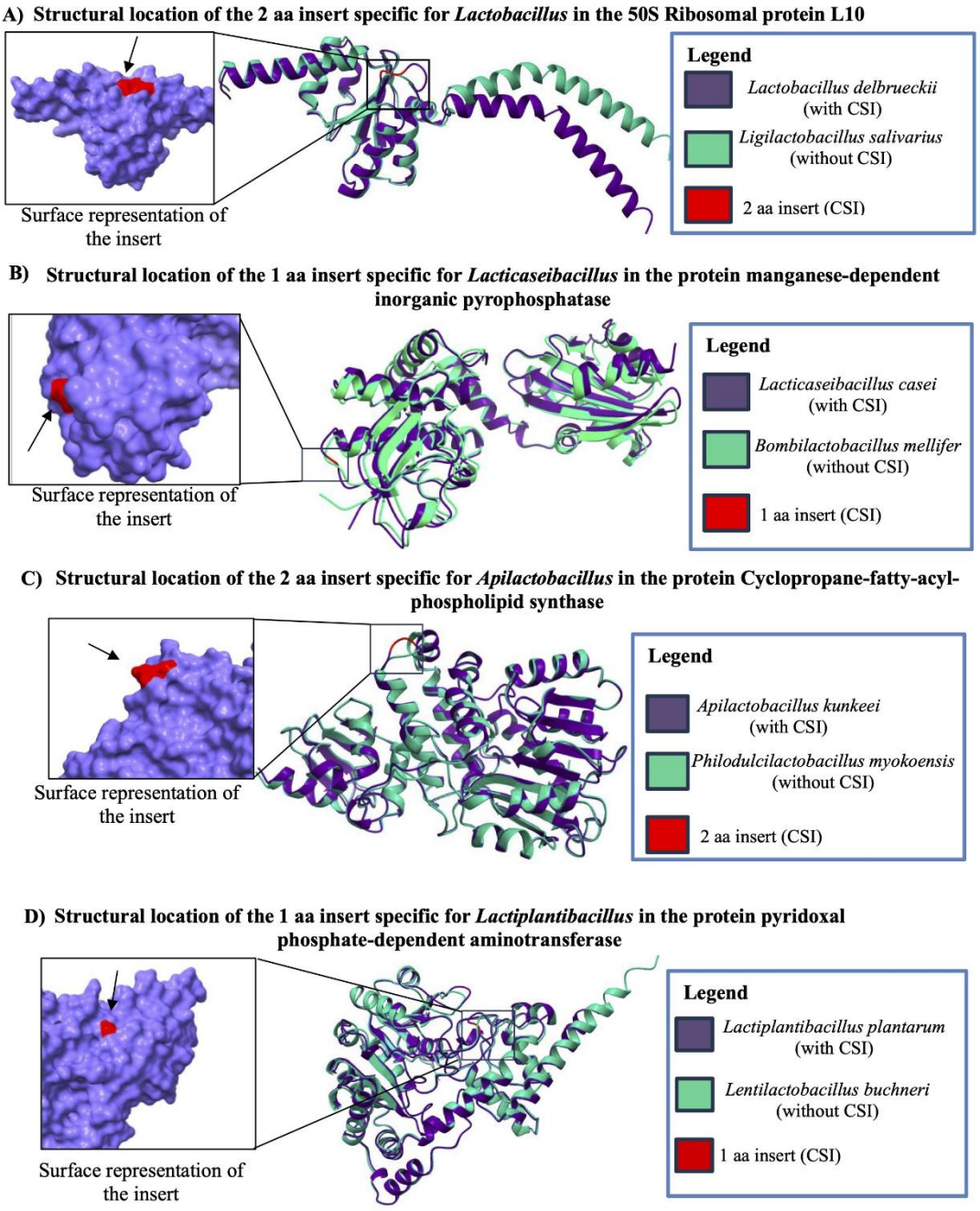

Figure 8. Superimposed cartoon and surface representations of AlphaFold-predicted protein structures showing CSIs specific to different *Lactobacillaceae* genera: (A) *Lactobacillus*-specific CSI in the 50S ribosomal protein L10 (RMSD = 5.4 Å); (B) *Lacticaseibacillus*-specific CSI in manganese-dependent inorganic pyrophosphatase (RMSD = 1.1 Å); (C) *Apilactobacillus*-specific CSI in cyclopropane-fatty-acyl-phospholipid synthase family protein (RMSD = 0.3 Å); and (D) *Lactiplantibacillus*-specific CSI in pyridoxal phosphate-dependent aminotransferase (RMSD = 1.0 Å). In each panel, the CSI-containing homolog is shown in dark purple, the CSI-lacking homolog in green, and the CSI position is highlighted in red. Further details on protein structure prediction and analysis are provided in the Methods section.

 doi:10.20944/preprints202506.2420.v2

We used AlphaFold [47] to predict the structures of proteins containing conserved signature inserts (CSIs), along with homologous proteins lacking these inserts. To determine the structural localization of each CSI, we superimposed the predicted structures of the protein homologs with and without the CSI. Figure 8 shows the results for four representative CSIs, with the CSI regions highlighted in red. In all cases, the CSIs were localized to surface-exposed loop regions of the proteins. The RMSD values for three of the studied proteins viz. manganese-dependent inorganic pyrophosphatase (Fig. 8B; RMSD = 1.1 Å), cyclopropane-fatty-acyl-phospholipid synthase (Fig. 8C; RMSD = 0.3 Å), and pyridoxal phosphate-dependent aminotransferase (Fig. 8D; RMSD = 1.0 Å), indicate minimal structural differences between the CSI-lacking and -containing proteins. In contrast, the 50S ribosomal protein L10, which contains a 2-amino acid insertion specific to *Lactobacillus* (Fig. 8A), showed a higher RMSD (5.4 Å), suggesting a potential conformational change induced by the CSI.

These findings on the CSIs specific for *Lactobacillaceae* genera are consistent with earlier studies that the conserved indels in protein sequences are structurally localized to surface loops [76,80,82,83], and may facilitate novel genus-specific functional interactions.

4. Discussion

Lactobacillaceae species play crucial roles in food production, probiotic development, and human health due to their metabolic versatility [1-3,5]. Given their importance, it is essential to understand the unique characteristics shared by species across different *Lactobacillaceae* genera. Traditional phylogenetic and genomic similarity-based approaches used for genus-level classification often fail to reveal genus-specific traits [3,15,32,84]. This study presents a comprehensive phylogenomic and comparative genomic analysis of 410 *Lactobacillaceae* genomes, leading to the identification of 167 novel molecular markers, termed CSIs or TAXIs, found in diverse proteins. Each CSI is uniquely associated with a specific *Lactobacillaceae* genus. The discovery of these genus-specific molecular markers represents a significant advancement in our understanding of the *Lactobacillaceae* family. These TAXIs not only enable more precise molecular delineation of genera [25,31], but also offer new tools for diagnostic development [85,86], and for genetic and biochemical studies aimed at uncovering genus-specific functional traits [78,80,82,87]. This enhances both scientific insight and practical applications of *Lactobacillaceae* species.

The predictive power of the identified TAXIs was demonstrated in this study through their successful use in classifying 111 out of 113 uncharacterized *Lactobacillus* isolates using the AppIndels.com server. This tool matches genome sequences against a curated database of known TAXIs, allowing rapid and accurate genus-level identification [37,71]. Such capabilities are especially valuable as genomic databases expand and new *Lactobacillaceae* species continue to be discovered. With the integration of TAXIs into the AppIndels.com server, it can also detect the presence of these genera in high-throughput genomic datasets and help resolve taxonomic ambiguities. For the genomes of two *Lactobacillus* isolates (GCA_014796685.1 and GCA_019303535.1) no taxonomic prediction was made by the server. The NCBI record of the genome GCA_014796685.1 indicates that it is contaminated, whereas the other genome (viz. GCA_019303535.1) could correspond to a *Lactobacillaceae* genus (viz. monotypic genera) for which no CSIs were identified in this study. While the AppIndels server provide a reliable tool for predicting taxonomic affiliation and supporting diagnostic applications based on genome sequences, its one key limitation is that it can only assign taxa to genera for which validated CSIs are present in its database [37]. Additionally, the server may fail to make a prediction or produce an incorrect result, if the genome sequence analyzed is contaminated or partial [37].

Beyond *in silico* detection, the identified TAXIs are also ideally suited for developing novel diagnostic assays. Because these CSIs are located within conserved regions of genes/proteins, sequences from their flanking regions can be used to design PCR primers or probes for qPCR and pyrosequencing, enabling selective amplification or detection of CSI-containing organisms [88]. Similar TAXI-based diagnostic assays have been successfully developed for other taxa, such as

 doi:10.20944/preprints202506.2420.v2

Bacillus anthracis [86] and *Escherichia coli* O157:H7 [85], demonstrating the broader applicability of this approach.

In addition to the utility of TAXIs for taxonomy and diagnostics studies, these taxon-specific molecular markers also provide a gateway to exploring genus-specific functional traits. Structural mapping of CSIs, including those identified in this study, shows that they are consistently located in surface-exposed loop regions of proteins [76,80,83]. These surface-exposed loops are flexible, solvent-accessible, and often involved in protein–protein or protein–ligand interactions [76,79]. Earlier experimental studies on selected CSIs have shown the functional importance of these CSIs for the group of organisms for which they are specific [89]. This suggests that the rare genetic changes represented by CSIs may underlie unique biochemical or phenotypic traits [76,82,83,89-92]. Therefore, further genetic and biochemical studies on the identified TAXIs could uncover novel metabolic or adaptive traits uniquely shared by species within individual genera.

It should be noted that although CSIs are powerful genus-specific molecular markers, determining their functional significance remains challenging. These indels are often small and located in conserved protein regions, making it difficult to directly link them to phenotypic traits or biochemical functions [79]. Many CSIs occur in proteins with poorly characterized roles, and their subtle structural effects may not lead to easily observable changes in cellular behavior [27,80]. To investigate the functional relevance of CSIs, several approaches can be employed. Structural modeling tools like AlphaFold can help localize CSIs within protein structures [47], while molecular dynamics simulations and docking studies may reveal how CSIs affect protein flexibility, stability, or binding interactions [78,80,82]. Experimental techniques such as site-directed mutagenesis and functional assays can validate the impact of specific indels [89,93,94]. Additionally, protein interaction studies and omics-based profiling (e.g., transcriptomics, proteomics) may uncover downstream effects [95,96]. Correlating the presence of a CSI with specific ecological or phenotypic traits of the species group [78,87] may also provide insights into its functional role.

In summary, this study identifies 167 genus-specific conserved signature indels (CSIs), or TAXIs, across *Lactobacillaceae* genomes, offering powerful molecular markers for taxonomic studies, diagnostic assay development, and functional trait discovery. Given the widespread industrial use of *Lactobacillaceae* species in food, health, and biotechnology, further biochemical and functional characterization of these TAXIs could uncover novel genus-specific traits with significant implications for microbial ecology and applied microbiology.

Supplementary Materials: The following data are available online at www.mdpi.com/link, Table S1. Summary of CSIs specific to species from the genera *Ligilactobacillus-Liquorilactobacillus* cluster, *Lapidilactobacillus*, *Bombilactobacillus*, and *Schleiferilactobacillus* Table S2. Summary of CSIs specific to species from the genera *Agrilactobacillus*, *Latilactobacillus*, *Loigolactobacillus*, and *Furfurilactobacillus*. Table S3. Summary of CSIs specific for species to the genera *Paucilactobacillus*, *Limosilactobacillus*, *Fructilactobacillus Lentilactobacillus*, and *Levilactobacillus*. Table S4. Summary of CSIs Specific to species from the genera, *Pediococcus*, *Companilactobacillus*, *Xylocopilactobacillus* and *Dellaglioa*. Table S5. Information on the genome sequences of 111 uncharacterized *Lactobacillus* isolates and their taxonomic affiliations predicted by the AppIndels.com server. Figure S1. An uncompressed version of the maximum likelihood tree shown in Figure 1 for the 410 genome-sequenced *Lactobacillaceae* species. The type species is indicated with a superscript "T". **Figures S2-S17:** Partial sequence alignments of the 50S ribosomal protein L10, excinuclease ABC subunit UvrC, Anaerobic ribonucleoside-triphosphate reductase, DNA-binding protein WhiA, Translation initiation factor IF-2, 50S ribosomal protein L4, TIGR01457 family HAD-type hydrolase, C69 family dipeptidase, YfbR-like 5'-deoxynucleotidase, class I SAM-dependent methyltransferase, Phosphate acyltransferase PlsX, DNA helicase PcrA, NADP-dependent phosphogluconate dehydrogenase, calcium-translocating P-type ATPase, ATP-binding protein, 16S rRNA (cytosine(1402)-N(4))-methyltransferase RsmH, showing CSIs/TAXIs that are specific for the genus *Lactobacillus*. **Figures S18-S26:** Partial sequence alignments of the manganese-dependent inorganic pyrophosphatase, hemolysin family protein, 1-acyl-sn-glycerol-3-phosphate acyltransferase, DUF1002 domain-containing protein, DeoR/GlpR family DNA-binding transcription regulator, DNA polymerase IV, YfcE family phosphodiesterase,

 doi:10.20944/preprints202506.2420.v2

methionine adenosyltransferase, showing CSIs/TAXIs that are specific for the genus *Lacticaseibacillus.* **Figures S27-S30:** Partial sequence alignments of the cyclopropane-fatty-acyl-phospholipid synthase family protein, DEAD/DEAH box helicase, Phosphate acetyltransferase, glucose-6-phosphate dehydrogenase, showing CSIs/TAXIs that are specific for the genus *Apilactobacillus*. **Figures S31-S38:** Partial sequence alignments of the pyridoxal phosphate-dependent aminotransferase, ABC transporter ATPase, acetyl-CoA carboxylase, 50S ribosomal protein L15, C69 family dipeptidase, GRP family sugar transporter, glycoside hydrolase family 13 protein, undecaprenyl-phosphate alpha-N-acetylglucosaminyl 1-phosphate transferase, showing CSIs/TAXIs that are specific for the genus *Lactiplantibacillus*. **Figures S39-S45.** Partial sequence alignments of the PolC-type DNA polymerase III protein, heat-inducible transcriptional repressor HrcA protein, the transcription termination/antitermination protein NusG, UDP-N-acetylmuramoyl-L-alanine-D-glutamate ligase protein, dihydroorotate dehydrogenase protein, tRNA dihydrouridine synthase B protein, DNA repair protein RecN protein, showing CSIs/TAXIs that are specific for the genus *Ligilactobacillus* and *Liquorilactobacillus*. **Figures S46-S49:** Partial sequence alignments of the PolC-type DNA polymerase III protein, type 2 isopentenyl-diphosphate Delta-isomerase protein, polysaccharide biosynthesis protein, hydroxymethylglutaryl-CoA synthase protein, showing CSIs/TAXIs that are specific for the genus *Lapidilactobacillus*. **Figures S50-S56:** Partial sequence alignments of the protein SMC-Scp complex subunit SepB, protein response regulator transcription factor, protein Translation initiation factor IF-3, protein YqeG family HAD IIIA-type phosphatase, protein response regulator transcription factor, protein arginine-tRNA ligase, protein DNA polymerase III subunit alpha, showing CSIs/TAXIs that are specific for the genus *Bombilactobacillus.* **Figures S57- S71:** Partial sequence alignments of the protein excinuclease ABC subunit UvrC, 50S ribosomal protein L15, response regulator transcription factor protein, HD domain-containing protein, metallophosphoesterase protein, 1,4-dihydroxy-2-naphthoate polyprenyltransferase protein, DNA polymerase III subunit gamma/tau protein, UDP-N-acetylmuramoyl-L-alanine-D-glutamate ligase protein, ABC-F family ATP-binding cassette domain-containing protein, glutamine-hydrolyzing GMP synthase protein, phosphopentomutase protein, citrate lyase acyl carrier protein, orotidine-5'-phosphate decarboxylase protein, protein molecular chaperone DnaK, protein oligoendopeptidase F, showing CSIs/TAXIs that are specific for the genus *Schleiferilactobacillus*. **Figures S72-S75:** Partial sequence alignments of the helix-turn-helix domain-containing protein, protein heat-inducible transcriptional repressor HrcA, MBL fold metallo-hydrolase protein, RluA family pseudouridine synthase protein showing CSIs/TAXIs that are specific for the genus *Agrilactobacillus*. **Figures S76-S83:** Partial sequence alignments of the protein MDR family MFS transporter, protein tRNA (adenine(22)-N(1))-methyltransferase TrmK, alanine racemase protein, competence protein ComEA, RNA polymerase recycling motor HelD, Two-component system regulatory protein YycI, nucleobase.cation symporter-2 family protein, protein calcium ABC transporter ATPase, showing CSIs/TAXIs that are specific for the genus *Latilactobacillus.* **Figures S84-S89:** Partial sequence alignments of the protein phosphoglycerate dehydrogenase, protein cation-translocating P-type ATPase, preprotein translocase subunit SecA, pyruvate carboxylase protein, amino acid permease, protein SAM-dependent methyltransferase, showing CSIs/TAXIs that are specific for the genus *Loigolactobacillus*. **Figures S90-S108**: Partial sequence alignments of the protein, GTPase ObgE, DHH family phosphoesterase, phosphate acyltransferase PlsX, CtsR family transcriptional regulator, energy-coupling factor ABC transporter ATP-binding protein, phosphoglycerate kinase, Nramp family divalent metal transporter protein, polysaccharide biosynthesis protein, peptidylprolyl isomerase, ribonuclease J, DNA-formamidopyrimidine glycosylase, class I mannose-6-phosphate isomerase, mechanosensitive ion channel family protein,dihydrolipoyl dehydrogenase, glutamate-tRNA ligase, LTA synthase family protein, PolC-type DNA polymerase III protein, peptidase M13 protein , showing CSIs/TAXIs that are specific for the genus *Furfurilactobacillus*. **Figures S109-S111:** Partial sequence alignments of the protein iron-containing alcohol dehydrogenase protein, protein phosphogluconate dehydrogenase (NAD(+)-dependent, decarboxylating), ROK family glucokinase protein, showing CSIs/TAXIs that are specific for the genus *Paucilactobacillus.* **Figures S112-S119**: Partial sequence alignments of the protein UTP-glucose-1-phosphate uridylyltransferase GalU, proline-tRNA ligase, S1-like domain-containing RNA-binding protein, (d)CMP kinase protein, Ammonia-dependent NAD(+) synthetase protein, tRNA uracil-4-sulfurtransferase ThiI protein, redox-regulated ATPase YchF protein, class I SAM-dependent RNA methyltransferase protein showing CSIs/TAXIs that are specific for the genus *Limosilactobacillus*. **Figures S120-S127:** Partial sequence alignments of DNA repair protein RecN, undecaprenyldiphospho-

 doi:10.20944/preprints202506.2420.v2

muramoylpentapeptide beta-N- protein, ribulose-phosphate 3-epimerase, nucleoside hydrolase, zinc-dependent alcohol dehydrogenase family protein, PBPLA family penicillin-binding protein, ribonuclease J, DNA topoisomerase (ATP-hydrolyzing) subunit B protein, showing CSIs/TAXIs that are specific for the genus *Fructilactobacillus*. **Figures S128-S130:** Partial sequence alignments of AI-ZE family transporter protein, endonuclease Mut S2, heat-inducible transcriptional proteinshowing CSIs/TAXIs that are specific for the genus *Lentilactobacillus*. **Figures S131-S134:** Partial sequence alignments of ATP-dependent DNA helicase RecG, pyridoxal phosphate-dependent aminotransferase, iron-sulfur cluster biosynthesis protein, EamA family transporter protein, showing CSIs/TAXIs that are specific for the genus *Levilactobacillus.* **Figures S135-S138:** Partial sequence alignments of 50S ribosomal protein L15, LacI family DNA-binding transcriptional regulator, trypsin-like peptidase domain-containing protein, BCCT family transporter protein, showing CSIs/TAXIs that are specific for the genus *Secundilactobacillus.* **Figures S139-S148:** Partial sequence alignments of 6 phosphofructokinase, glutamine-fructose-6-phosphate transaminase, ATP-dependent chaperone ClpB protein, endolytic transglycosylase MltG, PBP1A family penicillin-binding protein, cell division protein FtsA, histidine phosphatase family protein, proline-specific peptidase family protein, cyclopropane-fatty-acyl-phospholipid synthase family protein, aminopeptidase C, showing CSIs/TAXIs that are specific for the genus *Pediococcus*. **Figures S149-158:** Partial sequence alignments of SkL family PASTA domain-containing Ser/Thr kinase, type I glyceraldehyde-3-phosphate dehydrogenase, DNA polymerase III subunit beta, IMP dehydrogenase, cysteine-tRNA ligase, L-threonylcarbamoyladenylate synthase, ribonuclease J, ABC-F family ATP-binding cassette domain-containing protein, DNA-directed RNA polymerase subunit betashowing CSIs/TAXIs that are specific for the genus *Companilactobacillus*. **Figures S159-S162:** Partial sequence alignment of DNA polymerase III subunit alpha, excinuclease ABC subunit UvrC, ribosome biogenesis GTP-binding protein YihA/YsxC, tRNA uracil 4-sulfurtransferase ThiI, showing CSIs/TAXIs that are specific for the genus *Xylocopilactobacillus.* **Figures S163-S168:** Partial sequence alignment of DEAD/DEAH box helicase, amino acid ABC transporter substrate-binding protein/permease, FtsW/RodA/SpoVE family cell cycle protein, transglycosylase domain-containing protein, RNA polymerase sigma factor RpoD protein, amino acid ABC transporter permease protein, showing CSIs/TAXIs that are specific for the genus *Dellaglioa*. **Figure S169.** The results from AppIndels server showing predicted taxonomic affiliations for the genomes of two unclassified *Lactobacillus* isolates. (A) The *Lactobacillus* strain CBA3605 identified by the server as belonging to the genus *Lactiplantibacillus*,. (B) The genome of *Lactobacillus* strain UW_DM_LACCAS1_1 is predicted to be affiliated with the genus *Lacticaseibacillus*.

Author Contributions: SB and RSG carried out analysis using the AppIndels server; SB constructed phylogenetic trees; RSG, Planning and supervision of the work, obtained funding for the project and writing and finalizing of the manuscript; SB, updating the sequence information for the CSIs and checking and formatting different Figures and Tables, RSG and SB, writing and finalizing of the manuscript.

Acknowledgments: This work was supported by a by the research grant (RGPIN-2019-06397) from the Natural Science and Engineering Research Council of Canada and a grant support from the Ontario Research Fund.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Skerman, V.B.D.; McGowan, V.; Sneath, P.H.A. Approved lists of bacterial names. *Int J Syst Bacteriol* **1980**, *30*, 225-420.
2. Winslow, C.; Broadhurst, J.B., RE.; Krumwiede, C.R., LA, Smith GH. The Families and Genera of the Bacteria: Preliminary Report of the Committee of the Society of American Bacteriologists on Characterization and Classification of Bacterial Types. . *J. Bacteriol. Res.* **1917**, *2*, 505-566, doi:doi:10.1128/jb.2.5.505-566.1917.
3. Zheng, J.; Wittouck, S.; Salvetti, E.; Franz, C.M.A.P.; Harris, H.M.B.; Mattarelli, P.; O'Toole, P.W.; Pot, B.; Vandamme, P.; Walter, J.; et al. A taxonomic note on the genus *Lactobacillus:* Description of 23 novel genera, emended description of the genus *Lactobacillus* Beijerinck 1901, and union of *Lactobacillaceae* and *Leuconostocaceae*. *Int J Syst Evol Microbiol* **2020**, *70*, 2782-2858, doi:https://doi.org/10.1099/ijsem.0.004107.
4. Schleifer, K.H. Family V. *Leuconostocaceae* fam. nov. *Bergey's Manual of Systematic Bacteriology (The Firmicutes), 2nd edn*, **2009**, *3*, 624, doi:https://doi.org/10.1002/9781118960608.fbm00127.

5. Nieminen, T.T.S., E,; Endo, A.; Johansson, P.,; Bjorkroth,J. The family *Leuconostocacea*e In *The Prokaryotes: Firmicutes and Tenericutes:*; Springer-Verlag: 2014; pp. 215-240.
6. Parte, A.C.; Sarda Carbasse, J.; Meier-Kolthoff, J.P.; Reimer, L.C.; Goker, M. List of Prokaryotic names with Standing in Nomenclature (LPSN) moves to the DSMZ. *Int J Syst Evol Microbiol* **2020**, *70*, 5607-5612, doi:10.1099/ijsem.0.004332.
7. Danza, A.; Lucera, A.; Lavermicocca, P.; Lonigro, S.L.; Bavaro, A.R.; Mentana, A.; Centonze, D.; Conte, A.; Del Nobile, M.A. Tuna Burgers Preserved by the Selected *Lactobacillus paracasei* IMPC 4.1 Strain. *Food Bioproc Tech* **2018**, *11*, 1651-1661, doi:10.1007/s11947-018-2129-y.
8. Chen, Y.; Yu, L.; Qiao, N.; Xiao, Y.; Tian, F.; Zhao, J.; Zhang, H.; Chen, W.; Zhai, Q. *Latilactobacillus curvatus*: A Candidate Probiotic with Excellent Fermentation Properties and Health Benefits. *Foods* **2020**, *9*, doi:10.3390/foods9101366.
9. Liang, J.R.; Deng, H.; Hu, C.Y.; Zhao, P.T.; Meng, Y.H. Vitality, fermentation, aroma profile, and digestive tolerance of the newly selected *Lactiplantibacillus plantarum* and *Lacticaseibacillus paracasei* in fermented apple juice. *Front. Nutr.* **2022**, *9*, doi:10.3389/fnut.2022.1045347.
10. Dicks, L.; Endo, A. The Family *Lactobacillaceae*: Genera Other than *Lactobacillus*. In *The Prokaryotes: Firmicutes and Tenericutes*, Rosenberg, E., DeLong, E.F., Lory, S., Stackebrandt, E., Thompson, F., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2014; pp. 203-212.
11. Qiao, N.; Wittouck, S.; Mattarelli, P.; Zheng, J.; Lebeer, S.; Felis, G.E.; Gänzle, M.G. After the storm-Perspectives on the taxonomy of *Lactobacillaceae*. *JDS Commun* **2022**, *3*, 222-227, doi:10.3168/jdsc.2021-0183.
12. Salvetti, E.; Harris, H.M.B.; Felis, G.E.; O'Toole, P.W. Comparative Genomics of the Genus *Lactobacillus* Reveals Robust Phylogroups That Provide the Basis for Reclassification. *Appl. Environ. Microbiol.* **2018**, *84*, e00993-00918, doi:doi:10.1128/AEM.00993-18.
13. Wittouck, S.; Wuyts, S.; Meehan, C.J.; Noort, V.v.; Lebeer, S. A Genome-Based Species Taxonomy of the *Lactobacillus* Genus Complex. *mSystems* **2019**, *4*, e00264-00219, doi:doi:10.1128/mSystems.00264-19.
14. Duar, R.M.; Lin, X.B.; Zheng, J.; Martino, M.E.; Grenier, T.; Pérez-Muñoz, M.E.; Leulier, F.; Gänzle, M.; Walter, J. Lifestyles in transition: evolution and natural history of the genus *Lactobacillus*. *FEMS Microbiol Rev* **2017**, *41*, S27-s48, doi:10.1093/femsre/fux030.
15. Parks, D.H.; Chuvochina, M.; Waite, D.W.; Rinke, C.; Skarshewski, A.; Chaumeil, P.A.; Hugenholtz, P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol* **2018**, *36*, 996-1004.
16. Sun, Z.; Harris, H.M.; McCann, A.; Guo, C.; Argimón, S.; Zhang, W.; Yang, X.; Jeffery, I.B.; Cooney, J.C.; Kagawa, T.F.; et al. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. *Nat Commun* **2015**, *6*, 8322, doi:10.1038/ncomms9322.
17. Giraffa, G.; Chanishvili, N.; Widyastuti, Y. Importance of lactobacilli in food and feed biotechnology. *Research in Microbiology* **2010**, *161*, 480-487, doi:https://doi.org/10.1016/j.resmic.2010.03.001.
18. Zhang, Z.; Lv, J.; Pan, L.; Zhang, Y. Roles and applications of probiotic Lactobacillus strains. *Applied Microbiology and Biotechnology* **2018**, *102*, 8135-8143, doi:10.1007/s00253-018-9217-9.
19. Endo, A.; Maeno, S.; Tanizawa, Y.; Kneifel, W.; Arita, M.; Dicks, L.; Salminen, S. Fructophilic Lactic Acid Bacteria, a Unique Group of Fructose-Fermenting Microbes. *Appl Environ Microbiol* **2018**, *84*, doi:10.1128/aem.01290-18.
20. Sayers, E.W.; Agarwala, R.; Bolton, E.E.; Brister, J.R.; Canese, K.; Clark, K.; Connor, R.; Fiorini, N.; Funk, K.; Hefferon, T. Database resources of the national center for biotechnology information. *Nucleic Acids Res* **2019**, *47*, D23.
21. Parte, A.C. LPSN–List of Prokaryotic names with Standing in Nomenclature (bacterio. net), 20 years on. *Int J Syst Evol Microbiol* **2018**, *68*, 1825-1829.
22. Wu, L.; McCluskey, K.; Desmeth, P.; Liu, S.; Hideaki, S.; Yin, Y.; Moriya, O.; Itoh, T.; Kim, C.Y.; Lee, J.S.; et al. The global catalogue of microorganisms 10K type strain sequencing project: closing the genomic gaps for the validly published prokaryotic and fungi species. *Gigascience* **2018**, *7*, doi:10.1093/gigascience/giy026.
23. Mukherjee, S.; Seshadri, R.; Varghese, N.J.; Eloe-Fadrosh, E.A.; Meier-Kolthoff, J.P.; Göker, M.; Coates, R.C.; Hadjithomas, M.; Pavlopoulos, G.A.; Paez-Espino, D.; et al. 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat Biotechnol* **2017**, *35*, 676-683, doi:10.1038/nbt.3886.

24. Whitman, W.B. Genome sequences as the type material for taxonomic descriptions of prokaryotes 1. *Syst Appl Microbiol* **2015**, *38*, 217-222.
25. Bello, S.; Rudra, B.; Gupta, R.S. Phylogenomic and comparative genomic analyses of *Leuconostocaceae* species: identification of molecular signatures specific for the genera *Leuconostoc, Fructobacillus* and *Oenococcus* and proposal for a novel genus *Periweissella* gen. nov. *Int J Syst Evol Microbiol* **2022**, *72*, doi:https://doi.org/10.1099/ijsem.0.005284.
26. Gupta, R.S. Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiol Mol Biol Rev* **1998**, *62*, 1435-1491, doi:10.1128/mmbr.62.4.1435-1491.1998.
27. Gupta, R.S. Impact of genomics on the understanding of microbial evolution and classification: the importance of Darwin's views on classification. *FEMS Microbiol Rev* **2016**, *40*, 520-553, doi:10.1093/femsre/fuw011.
28. Rokas, A.; Holland, P.W.H. Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.* **2000**, *15*, 454-459, doi:https://doi.org/10.1016/S0169-5347(00)01967-4.
29. Bhandari, V.; Naushad, H.S.; Gupta, R.S. Protein based molecular markers provide reliable means to understand prokaryotic phylogeny and support Darwinian mode of evolution. *Front Cell Infect Microbiol* **2012**, *2*, 98, doi:10.3389/fcimb.2012.00098.
30. Naushad, H.S.; Lee, B.; Gupta, R.S. Conserved signature indels and signature proteins as novel tools for understanding microbial phylogeny and systematics: identification of molecular signatures that are specific for the phytopathogenic genera *Dickeya, Pectobacterium* and *Brenneria*. *Int J Syst Evol Microbiol* **2014**, *64*, 366-383, doi:10.1099/ijs.0.054213-0.
31. Gupta, R.S.; Patel, S.; Saini, N.; Chen, S. Robust demarcation of 17 distinct *Bacillus* species clades, proposed as novel *Bacillaceae* genera, by phylogenomics and comparative genomic analyses: description of *Robertmurraya kyonggiensis* sp. nov. and proposal for an emended genus *Bacillus* limiting it only to the members of the Subtilis and Cereus clades of species. *Int J Syst Evol Microbiol* **2020**, *70*, 5753-5798, doi:https://doi.org/10.1099/ijsem.0.004475.
32. Konstantinidis, K.T.; Tiedje, J.M. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol* **2007**, *10*, 504-509, doi:10.1016/j.mib.2007.08.006.
33. Moreira, D.; Philippe, H. Molecular phylogeny: pitfalls and progress. *Int. Microbiol* **2000**, *3*, 9-16.
34. Baldauf, S.L. Phylogeny for the faint of heart: a tutorial. *Trends Genet* **2003**, *19*, 345-351, doi:10.1016/s0168-9525(03)00112-4.
35. Gupta, R.S. Identification of Conserved Indels that are Useful for Classification and Evolutionary Studies. In *Bacterial Taxonomy, Methods in microbiology*, Goodfellow, M., Sutcliffe, I., Chun, J., Eds.; Elsevier: 2014; Volume 41, pp. 153-182.
36. Gupta, R.S.; Rudra, B.; Tony, J.; Bello, S. Phylogenomic and comparative analyses on protein sequences from Halobacteria to identify taxon-specific molecular markers which demarcate different *Halobacteriaceae* and *Haloarculaceae* genera. *Int J Syst Evol Microbiol* **2025**, *75*, doi:10.1099/ijsem.0.006879.
37. Gupta, R.S.; Kanter-Eivin, D. AppIndels.com Server: A Web Based Tool for the Identification of Known Taxon-Specific Conserved Signature Indels in Genome Sequences: Validation of Its Usefulness by Predicting the Taxonomic Affiliation of >700 Unclassified strains of *Bacillus* Species. *Int J Syst Evol Microbiol* **2023**, *73*.
38. Xiao, Y.; Zhao, J.; Zhang, H.; Zhai, Q.; Chen, W. Mining genome traits that determine the different gut colonization potential of Lactobacillus and Bifidobacterium species. *Microb Genom* **2021**, *7*, doi:10.1099/mgen.0.000581.
39. Wang, Z.; Wu, M. A phylum-level bacterial phylogenetic marker database. *Mol Biol Evol* **2013**, *30*, 1258-1262, doi:10.1093/molbev/mst059.
40. Adeolu, M.; Alnajar, S.; Naushad, S.; R, S.G. Genome-based phylogeny and taxonomy of the '*Enterobacteriales*': proposal for *Enterobacterales* ord. nov. divided into the families *Enterobacteriaceae, Erwiniaceae* fam. nov., *Pectobacteriaceae* fam. nov., *Yersiniaceae* fam. nov., *Hafniaceae* fam. nov., *Morganellaceae* fam. nov., and *Budviciaceae* fam. nov. *Int J Syst Evol Microbiol* **2016**, *66*, 5575-5599, doi:10.1099/ijsem.0.001485.

41. Eddy, S.R. Profile hidden Markov models. *Bioinformatics* **1998**, *14*, 755-763, doi:10.1093/bioinformatics/14.9.755.
42. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539, doi:https://doi.org/10.1038/msb.2011.75.
43. Capella-Gutiérrez, S.; Silla-Martínez, J.M.; Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **2009**, *25*, 1972-1973, doi:10.1093/bioinformatics/btp348.
44. Whelan, S.; Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **2001**, *18*, 691-699, doi:10.1093/oxfordjournals.molbev.a003851.
45. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **2018**, *35*, 1547-1549, doi:10.1093/molbev/msy096.
46. Jeanmougin, F.; Thompson, J.D.; Gouy, M.; Higgins, D.G.; Gibson, T.J. Multiple sequence alignment with Clustal X. *Trends Biochem Sci* **1998**, *23*, 403-405, doi:10.1016/s0968-0004(98)01285-7.
47. Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A.J.; Bambrick, J.; et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493-500, doi:10.1038/s41586-024-07487-w.
48. Schrödinger, L. The PyMOL Molecular Graphics System, Version 1.8. *(No Title)* **2015**.
49. Mariani, V.; Biasini, M.; Barbato, A.; Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **2013**, *29*, 2722-2728, doi:10.1093/bioinformatics/btt473.
50. Zhang, Y.; Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 702-710, doi:https://doi.org/10.1002/prot.20264.
51. Guo, H.B.; Perminov, A.; Bekele, S.; Kedziora, G.; Farajollahi, S.; Varaljay, V.; Hinkle, K.; Molinero, V.; Meister, K.; Hung, C.; et al. AlphaFold2 models indicate that protein sequence determines both structure and dynamics. *Sci Rep* **2022**, *12*, 10696, doi:10.1038/s41598-022-14382-9.
52. Dobritsa, A.P.; Linardopoulou, E.V.; Samadpour, M. Transfer of 13 species of the genus *Burkholderia* to the genus *Caballeronia* and reclassification of *Burkholderia jirisanensis* as *Paraburkholderia jirisanensis* comb. nov. *Int J Syst Evol Microbiol* **2017**, *67*, 3846-3853, doi:10.1099/ijsem.0.002202.
53. Montecillo, J.A.V.; Bae, H. Reclassification of *Brevibacterium frigoritolerans* as *Peribacillus frigoritolerans* comb. nov. based on phylogenomics and multiple molecular synapomorphies. *Int J Syst Evol Microbiol* **2022**, *72*, doi:10.1099/ijsem.0.005389.
54. Jiang, L.; Wang, D.; Kim, J.-S.; Lee, J.H.; Kim, D.-H.; Kim, S.W.; Lee, J. Reclassification of genus *Izhakiella* into the family *Erwiniaceae* based on phylogenetic and genomic analyses. *Int J Syst Evol Microbiol* **2021**, *70*, 3541-3546.
55. Gänzle, M.G. Lactic metabolism revisited: metabolism of lactic acid bacteria in food fermentations and food spoilage. *Curr. Opin. Food Sci.* **2015**, *2*, 106-117, doi:https://doi.org/10.1016/j.cofs.2015.03.001.
56. Zheng, J.; Ruan, L.; Sun, M.; Gänzle, M. A Genomic View of Lactobacilli and Pediococci Demonstrates that Phylogeny Matches Ecology and Physiology. *Appl Environ Microbiol* **2015**, *81*, 7233-7243, doi:10.1128/aem.02116-15.
57. Collins, M.D.; Rodrigues, U.; Ash, C.; Aguirre, M.; Farrow, J.A.E.; Martinez-Murcia, A.; Phillips, B.A.; Williams, A.M.; Wallbanks, S. Phylogenetic analysis of the genus *Lactobacillus* and related lactic acid bacteria as determined by reverse transcriptase sequencing of 16S rRNA. *FEMS Microbiol Lett* **1991**, *77*, 5-12, doi:https://doi.org/10.1111/j.1574-6968.1991.tb04313.x.
58. Salvetti, E.; Torriani, S.; Felis, G.E. The Genus *Lactobacillus:* A Taxonomic Update. *Probiotics and Antimicrobial Proteins* **2012**, *4*, 217-226, doi:10.1007/s12602-012-9117-8.
59. Shah, A.B.; Baiseitova, A.; Zahoor, M.; Ahmad, I.; Ikram, M.; Bakhsh, A.; Shah, M.A.; Ali, I.; Idress, M.; Ullah, R.; et al. Probiotic significance of Lactobacillus strains: a comprehensive review on health impacts, research gaps, and future prospects. *Gut Microbes* **2024**, *16*, 2431643, doi:10.1080/19490976.2024.2431643.
60. van de Guchte, M.; Penaud, S.; Grimaldi, C.; Barbe, V.; Bryson, K.; Nicolas, P.; Robert, C.; Oztas, S.; Mangenot, S.; Couloux, A.; et al. The complete genome sequence of Lactobacillus bulgaricus reveals

 doi:10.20944/preprints202506.2420.v2

extensive and ongoing reductive evolution. *Proc Natl Acad Sci U S A* **2006**, *103*, 9274-9279, doi:10.1073/pnas.0603024103.

61. Dan, T.; Hu, H.; Tian, J.; He, B.; Tai, J.; He, Y. Influence of Different Ratios of Lactobacillus delbrueckii subsp. bulgaricus and Streptococcus thermophilus on Fermentation Characteristics of Yogurt. *Molecules* **2023**, *28*, 2123.
62. Diaconu, M.; Kothe, U.; Schlünzen, F.; Fischer, N.; Harms, J.M.; Tonevitsky, A.G.; Stark, H.; Rodnina, M.V.; Wahl, M.C. Structural basis for the function of the ribosomal L7/12 stalk in factor binding and GTPase activation. *Cell* **2005**, *121*, 991-1004, doi:10.1016/j.cell.2005.04.015.
63. Lai, W.-K.; Lu, Y.-C.; Hsieh, C.-R.; Wei, C.-K.; Tsai, Y.-H.; Chang, F.-R.; Chan, Y. Developing Lactic Acid Bacteria as an Oral Healthy Food. *Life* **2021**, *11*, 268.
64. De Groote, M.A.; Frank, D.N.; Dowell, E.; Glode, M.P.; Pace, N.R. *Lactobacillus rhamnosus* GG bacteremia associated with probiotic use in a child with short gut syndrome. *Pediatr Infect Dis J* **2005**, *24*, 278-280, doi:10.1097/01.inf.0000154588.79356.e6.
65. Maeno, S.; Nishimura, H.; Tanizawa, Y.; Dicks, L.; Arita, M.; Endo, A. Unique niche-specific adaptation of fructophilic lactic acid bacteria and proposal of three *Apilactobacillus* species as novel members of the group. *BMC Microbiology* **2021**, *21*, 41, doi:10.1186/s12866-021-02101-9.
66. Bradford, E.L.; Wax, N.; Bueren, E.K.; Walke, J.B.; Fell, R.; Belden, L.K.; Haak, D.C. Comparative genomics of *Lactobacillaceae* from the gut of honey bees, *Apis mellifera,* from the Eastern United States. *G3 (Bethesda)* **2022**, *12*, doi:10.1093/g3journal/jkac286.
67. Fidanza, M.; Panigrahi, P.; Kollmann, T.R. Lactiplantibacillus plantarum-Nomad and Ideal Probiotic. *Front Microbiol* **2021**, *12*, 712236, doi:10.3389/fmicb.2021.712236.
68. Wang, Z.; Wu, J.; Tian, Z.; Si, Y.; Chen, H.; Gan, J. The Mechanisms of the Potential Probiotic *Lactiplantibacillus plantarum* against Cardiovascular Disease and the Recent Developments in its Fermented Foods. *Foods* **2022**, *11*, doi:10.3390/foods11172549.
69. Tohno, M.; Tanizawa, Y.; Sawada, H.; Sakamoto, M.; Ohkuma, M.; Kobayashi, H. A novel species of lactic acid bacteria, *Ligilactobacillus pabuli* sp. nov., isolated from alfalfa silage. *Int J Syst Evol Microbiol* **2022**, *72*, doi:https://doi.org/10.1099/ijsem.0.005587.
70. Barbour, A.G.; Adeolu, M.; Gupta, R.S. Division of the genus *Borrelia* into two genera (corresponding to Lyme disease and relapsing fever groups) reflects their genetic and phenotypic distinctiveness and will lead to a better understanding of these two groups of microbes (Margos et al. (2016) There is inadequate evidence to support the division of the genus *Borrelia*. Int. J. Syst. Evol. Microbiol. doi: 10.1099/ijsem.0.001717). *Int J Syst Evol Microbiol* **2017**, *67*, 2058-2067.
71. Rudra, B.; Gupta, R.S. Molecular Markers Specific for the *Pseudomonadaceae* Genera Provide Novel and Reliable Means for the Identification of Other *Pseudomonas* strains/spp. Related to These Genera. *Genes* **2025**, *16*, 183, doi:https://doi.org/10.3390/genes16020183.
72. Malhotra, M.; Bello, S.; Gupta, R.S. Phylogenomic and molecular markers based studies on clarifying the evolutionary relationships among *Peptoniphilus* species. Identification of several Genus-Level clades of *Peptoniphilus* species and transfer of some *Peptoniphilus* species to the genus *Aedoeadaptatus*. *Syst Appl Microbiol* **2024**, *47*, 126499, doi:10.1016/j.syapm.2024.126499.
73. Fusco, V.; Chieffi, D.; Fanelli, F.; Montemurro, M.; Rizzello, C.G.; Franz, C.M. The Weissella and Periweissella genera: Up-to-date taxonomy, ecology, safety, biotechnological, and probiotic potential. *Frontiers in Microbiology* **2023**, *14*, 1289937.
74. Endo, A.; Okada, S. Reclassification of the genus Leuconostoc and proposals of Fructobacillus fructosus gen. nov., comb. nov., Fructobacillus durionis comb. nov., Fructobacillus ficulneus comb. nov. and Fructobacillus pseudoficulneus comb. nov. *Int J Syst Evol Microbiol* **2008**, *58*, 2195-2205, doi:10.1099/ijs.0.65609-0.
75. Gupta, R.S.K.-E., D. AppIndels.com Server: A Web Based Tool for the Identification of Known Taxon-Specific Conserved Signature Indels in Genome Sequences: Validation of Its Usefulness by Predicting the Taxonomic Affiliation of >700 Unclassified strains of Bacillus Species

Int J Syst Evol Microbiol **2023**, In press.

 doi:10.20944/preprints202506.2420.v2

76. Akiva, E.; Itzhaki, Z.; Margalit, H. Built-in loops allow versatility in domain-domain interactions: lessons from self-interacting domains. *Proc Natl Acad Sci U S A* **2008**, *105*, 13292-13297, doi:10.1073/pnas.0801207105.
77. Hormozdiari, F.; Salari, R.; Hsing, M.; Schönhuth, A.; Chan, S.K.; Sahinalp, S.C.; Cherkasov, A. The effect of insertions and deletions on wirings in protein-protein interaction networks: a large-scale study. *J Comput Biol* **2009**, *16*, 159-167, doi:10.1089/cmb.2008.03TT.
78. Khadka, B.; Persaud, D.; Gupta, R.S. Novel Sequence Feature of SecA Translocase Protein Unique to the Thermophilic Bacteria: Bioinformatics Analyses to Investigate Their Potential Roles. *Microorganisms* **2020**, *8*, 59.
79. Miton, C.M.; Tokuriki, N. Insertions and Deletions (Indels): A Missing Piece of the Protein Engineering Jigsaw. *Biochemistry* **2023**, *62*, 148-157, doi:10.1021/acs.biochem.2c00188.
80. Gupta, R.S.; Nanda, A.; Khadka, B. Novel molecular, structural and evolutionary characteristics of the phosphoketolases from *Bifidobacteria* and *Coriobacteriales*. *PLoS One* **2017**, *12*, e0172176, doi:10.1371/journal.pone.0172176.
81. Geszvain, K.; Gruber, T.M.; Mooney, R.A.; Gross, C.A.; Landick, R. A hydrophobic patch on the flap-tip helix of E.coli RNA polymerase mediates sigma(70) region 4 function. *J Mol Biol* **2004**, *343*, 569-587, doi:10.1016/j.jmb.2004.08.063.
82. Khadka, B.; Gupta, R.S. Identification of a conserved 8 aa insert in the PIP5K protein in the Saccharomycetaceae family of fungi and the molecular dynamics simulations and structural analysis to investigate its potential functional role. *Proteins* **2017**, *85*, 1454-1467, doi:10.1002/prot.25306.
83. Hashimoto, K.; Panchenko, A.R. Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proc Natl Acad Sci U S A* **2010**, *107*, 20352-20357, doi:10.1073/pnas.1012999107.
84. Qin, Q.L.; Xie, B.B.; Zhang, X.Y.; Chen, X.L.; Zhou, B.C.; Zhou, J.; Oren, A.; Zhang, Y.Z. A proposed genus boundary for the prokaryotes based on genomic insights. *J Bacteriol* **2014**, *196*, 2210-2215, doi:10.1128/jb.01688-14.
85. Wong, S.Y.; Paschos, A.; Gupta, R.S.; Schellhorn, H.E. Insertion/deletion-based approach for the detection of *Escherichia coli* O157:H7 in freshwater environments. *Environ Sci Technol* **2014**, *48*, 11462-11470, doi:10.1021/es502794h.
86. Ahmod, N.Z.; Gupta, R.S.; Shah, H.N. Identification of a *Bacillus anthracis* specific indel in the yeaC gene and development of a rapid pyrosequencing assay for distinguishing *B. anthracis* from the *B. cereus* group. *J Microbiol Methods* **2011**, *87*, 278-285, doi:10.1016/j.mimet.2011.08.015.
87. Hassan, F.M.N.; Gupta, R.S. Novel Sequence Features of DNA Repair Genes/Proteins from Deinococcus Species Implicated in Protection from Oxidatively Generated Damage. *Genes (Basel)* **2018**, *9*.
88. Griffiths, E.; Petrich, A.; Gupta, R.S. Conserved Indels in Essential Proteins that are Distinctive Characteristics of *Chlamydiales* and Provide Novel Means for Their Identification. *Microbiology* **2005**, *151*, 2647-2657.
89. Singh, B.; Gupta, R.S. Conserved inserts in the Hsp60 (GroEL) and Hsp70 (DnaK) proteins are essential for cellular growth. *Mol. Genet. Genomics* **2009**, *281*, 361-373.
90. Clarke, J.H.; Irvine, R.F. Evolutionarily conserved structural changes in phosphatidylinositol 5-phosphate 4-kinase (PI5P4K) isoforms are responsible for differences in enzyme activity and localization. *Biochem J* **2013**, *454*, 49-57, doi:10.1042/bj20130488.
91. Kuznedelov, K.; Minakhin, L.; Niedziela-Majka, A.; Dove, S.L.; Rogulja, D.; Nickels, B.E.; Hochschild, A.; Heyduk, T.; Severinov, K. A role for interaction of the RNA polymerase flap domain with the sigma subunit in promoter recognition. *Science* **2002**, *295*, 855-857, doi:10.1126/science.1066303.
92. Nandan, D.; Lopez, M.; Ban, F.; Huang, M.; Li, Y.; Reiner, N.E.; Cherkasov, A. Indel-based targeting of essential proteins in human pathogens that have close host orthologue(s): discovery of selective inhibitors for Leishmania donovani elongation factor-1alpha. *Proteins* **2007**, *67*, 53-64.
93. Zeng, F.; Zhang, Y.; Zhang, Z.; Malik, A.A.; Lin, Y. Multiple-site fragment deletion, insertion and substitution mutagenesis by modified overlap extension PCR. *Biotechnology & Biotechnological Equipment* **2017**, *31*, 339-348, doi:10.1080/13102818.2017.1279033.

 doi:10.20944/preprints202506.2420.v2

94. Richardson, R.M.; Pascal, S.M. Bacterial two-hybrid systems evolved: innovations for protein-protein interaction research. *Journal of Bacteriology* **2025**, *207*, e00129-00125, doi:doi:10.1128/jb.00129-25.
95. Kothe, J.A.F.; Sauerwein, T.; Dietz, T.; Scheuer, R.; Elhossary, M.; Barth-Weber, S.; Wähling, J.; Förstner, K.U.; Evguenieva-Hackenberg, E. Early posttranscriptional response to tetracycline exposure in a gram-negative soil bacterium reveals unexpected attenuation mechanism of a DUF1127 gene. *RNA Biol* **2025**, *22*, 1-16, doi:10.1080/15476286.2025.2521887.
96. Cordwell, S.J. Exploring and Exploiting Bacterial Proteomes. In *Genomics, Proteomics, and Clinical Bacteriology: Methods and Reviews*, Woodford, N., Johnson, A.P., Eds.; Humana Press: Totowa, NJ, 2004; pp. 115-135.