Comparative genomics of the lactic acid bacteria


Comparative genomics of the lactic acid bacteria (LAB) are associated with various plant and animal niches and play a key role in the production of fermented foods and beverages. We report nine genome sequences representing the phylogenetic and functional diversity of these bacteria. The small genomes of lactic acid bacteria encode a broad repertoire of transporters for efficient carbon and nitrogen acquisition from the nutritionally rich environments they inhabit and reflect a limited range of biosynthetic capabilities that indicate both prototrophic and auxotrophic strains. Phylogenetic analyses, comparison of gene content across the group, and reconstruction of ancestral gene sets indicate a combination of extensive gene loss and key gene acquisitions via horizontal gene transfer during the coevolution of lactic acid bacteria with their habitats.

General Features of the LAB Genomes. The major features of the sequenced LAB genomes are summarized in Table 1, which is published as supporting information on thePNAS web site. The number of predicted protein-coding genes in the LAB differs from ~1,700 to ~2,800. Given the close phylogenetic relationship of these organisms, such a difference suggests substantial gene loss and/or gain in their evolution. In addition, all LAB genomes harbor pseudogenes. Strikingly, the number of pseudogenes differs by an order of magnitude, from <20 in Leuconostoc mesenteroides and Pediococcus pentosaceus to ~200 in S. thermophilus and Lactobacillus delbrueckii, indicating an active, ongoing process of genome degeneration. The LAB also differ in the number of rRNA operons, from two in Oenococcus oeni to nine in Lb. delbrueckii, which correlates with the number of tRNA genes (Table 1) and may reflect differences in the ecological competitiveness (e.g., capacity for rapid growth and production of lactic acid) between these bacteria (12, 13).


The authors declare no conflict of interest.

Abbreviations: LAB, lactic acid bacteria; COG, cluster of orthologous genes; LaCOG, lactic acid bacteria-specific COG; HGT, horizontal gene transfer.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. CP000411–CP000430).

PNAS October 17, 2006 vol. 103 no. 42 15611–15616
LAB genomes contained transposons, ranging from \( \approx 0.2\% \) of the genome in *Lactobacillus gasseri* to nearly 5% in *L. lactis* ssp. *cremoris*. Many LAB harbor plasmids, some of which are essential for growth in specific environments and carry genes for metabolic pathways, membrane transport, and bacteriocin production (14). The contribution of plasmid-encoded genes in the LAB ranges from 0% to 4.8% of total gene content (Table 1).

**Phylogenetic Analysis and Impact of Horizontal Gene Transfer (HGT) on LAB Evolution.** The LAB analyzed here belong to the phylum Firmicutes, class Bacilli and order Lactobacillales, a sister taxon to Firmicutes, Bacilli, and Bacillales. Classification of Lactobacillales remains an unresolved issue in particular because phenotypic classification, which is traditionally based on the type of fermentation, does not match the rRNA-based phylogeny (15). Whole-genome DNA and DNA–RNA hybridization and GC content studies led to the delineation of three closely related lineages of Lactobacillales (16): the Leuconostoc group (L. mesenteroides and *O. oeni*), the Lactobacillus casei-Pediococcus group (Lb. plantarum, Lb. casei, P. pentosaceus, and Lactobacillus brevis), and the Lb. delbrueckii group (Lb. delbrueckii, Lb. gasseri, and Lb. johnsonii) (15); streptococci (S. thermophilus) and lactococci (Lc. lactis ssp. lactis and *L. lactis* ssp. *cremoris*) formed a separate branch (16).

Fig. 1. Phylogenetic trees of Lactobacillales constructed on the basis of concatenated alignments of ribosomal proteins. All branches are supported at \( >75\% \) bootstrap values. Species are colored according to the current taxonomy: Lactobacillaceae, blue; Leuconostocaceae, magenta; Streptococaceae, red.

The availability of complete genomes for all major branches of Lactobacillales enables a more definitive analysis of their evolutionary relationships. We constructed phylogenetic trees from concatenated protein sequences, an approach shown to improve the resolution and increase robustness of phylogenetic analyses (17). We supplemented the ribosomal protein data set (Fig. 1) with concatenated RNA polymerase subunits (Fig. 4), which is published as supporting information on the PNAS web site), which also undergo little horizontal transfer. Both trees, constructed with a variety of methods, display the same topology with strongly supported internal branches. The Streptococci-lactococci branch is basal in the Lactobacillales tree, and the Pediococcus group is a sister to the Leuconostoc group within the Lactobacillus clade. Thus, the Lactobacillus genus appears to be paraphyletic with respect to the Pediococcus–Leuconostoc group. Lactobacillus casei is confidently placed at the base of the Lb. delbrueckii group, which contradicts the previous classifications (16, 18).

A molecular clock test (19) showed a high heterogeneity of evolutionary rates within Lactobacillales. Most of the root-to-tip distances are significantly unequal to the mean tree height; the previously reported (20) accelerated evolution of the Leuconostoc group (by a factor of 1.7–1.9 relative to the sister Pediococcus group) was especially prominent.

The strength of purifying selection acting on Lactobacillales species can be estimated by using two closely related pairs of genomes: *Lb. gasseri*/*Lb. johnsonii* and *Lc. lactis*/*Lc. cremoris*. Synonymous and non-synonymous substitution rates were estimated from concatenated coding sequence alignments of 443 orthologous genes (142,031 codons). The ds/dN (distance at synonymous sites/distance at non-synonymous sites) ratio was 38.5 \( \pm 0.5 \) for the *Lb. gasseri*/*Lb. johnsonii* pair and 29.8 \( \pm 0.4 \) for *Lc. lactis*/*Lc. cremoris* pair, showing unusually strong evolutionary pressure as compared with Proteobacteria, which has a characteristic ds/dN ratio of 5–10 (21). This is likely to reflect the large effective population size and/or high mutation rate of the Lactobacillales species because the intensity of purifying selection is known to be proportional to these quantities (22).

**Clusters of Orthologous Genes in Lactobacillales.** Robust identification of sets of orthologs (genes derived from the same ancestral gene) is a prerequisite for informative evolutionary–genomic analysis of any group of organisms. By using the computational procedure described previously (23), we constructed Lactobacillales-specific clusters of orthologous genes [LaCOGs (abbreviated COGs if clusters of orthologous genes are not Lactobacillales-specific)] for the Lactobacillales-specific set (Table 2, which is published as supporting information on the PNAS web site) from proteins encoded in 12 sequenced Lactobacillales genomes that were available at the time of this analysis. Many COGs include paralogous genes that evolved via duplications at different stages of evolution. The construction of orthologous clusters for a compact taxon, such as the Lactobacillales, results in much finer granularity, with a greater fraction of clusters containing a single member from all or most of the analyzed species. Altogether, 3,199 LaCOGs, which included from 2 species to all 12 species, were identified. On average, LaCOGs covered 86% of the genome (Fig. 2); 1,133 (35%) LaCOGs showed a one-to-one correspondence with the general COG set; 1,359 (43%) LaCOGs corresponded to 390 COGs that have been split into two or more paralogous groups. The remaining 707 (22%) LaCOGs have no counterparts in the general COG set (24); of these, 338 (11%) were shared with one or more non-Lactobacillales bacterial genomes among those reported recently and not yet included in the COGs, and 369 (11%) appeared to be specific to the Lactobacillales. Thus, the LaCOGs are a powerful resource for genome annotation and evolutionary analysis of Lactobacillales (for details, see Table 2).

The conserved core of genes present in all 12 species analyzed in the Lactobacillales genomes (Table 2) consists of 567 LaCOGs (18%). Functional distribution of the LaCOGs in this core shows that the majority encode components of the information-processing systems (translation, transcription, and replication). However, the core also includes 41 uncharacterized genes and 50 genes with only a general prediction of biochemical activity. Because these genes are conserved throughout Lactobacillales, it is likely that they have essential functions, at least within this group. Furthermore, two core genes have no detectable orthologs outside lactobacilli. One of these unique genomic markers of Lactobacillales contains a LysM (peptidoglycan-binding) domain (LaCOG01826). In several lactobacilli, this gene is located next to the genes for ribosomal proteins and cytidylate kinase and might be coregulated with these housekeeping genes. The second genomic marker, the highly conserved LaCOG01237, contains no characterized domains. However, this gene is located in a conserved genomic neighborhood encoding two enzymes implicated in 4-thiouridine modification of tRNA [(5-methylaminomethyl-2-thiouridylyl) methyltransferase and...
a predicted sulfurase] (LaCOG00578 and LaCOG01188; see Table 2), suggesting a role of LaCOG01237 proteins in specific modulation of this essential modification (25).

**Local Molecular Clock and HGT.** We tested the consistency of a local molecular clock in individual LaCOGs with a technique developed recently (26). A matrix of interspecies distances for each LaCOG was compared with the matrix of baseline distances obtained from the concatenated alignment of ribosomal proteins. If a COG evolves in a clock-like manner relative to the local molecular clock (Table 3, which is published as supporting information on the PNAS web site). Several functional groups of genes show statistically significant differences in their propensity to violate the local molecular clock (Table 3, which is published as supporting information on the PNAS web site). The acquisition of the mevalonate pathway tree with a bacterial intermediate). This pathway displaced the ancestral bacterial deoxyxylulose pathway of isoprenoid biosynthesis (27). Noteworthy is the potential for such transfer in Lc. lactis ssp. cremoros SK11 that harbors a conjugative plasmid (pLAC3) and several additional plasmids encoding genes related to growth in milk (28).

**Reconstruction of Gene Gain and Loss in the Evolution of Lactobacillales.** We used a version of the weighted parsimony algorithm (29) to reconstruct the events that occurred during the evolution of this group after its divergence from the common ancestor of all Bacilli (see Materials and Methods). This reconstruction suggests that the common ancestor of Lactobacillales had at least \(\approx 2,100-2,200\) genes, losing 600-1,200 genes (\(\approx 25-30\%\)) and gaining \(<100\) genes compared with the ancestor of all Bacilli, for which the genome size of \(\approx 2,700-3,700\) genes is predicted (Fig. 3). Thus, the origin of Lactobacillales involved extensive loss of ancestral genes (Table 5, which is published as supporting information on the PNAS web site). Many of the changes mapped to this stage of evolution seem to be related to the transition to life in a nutritionally rich medium. Thus, a number of genes for biosynthesis of cofactors were lost; conversely, a variety of peptidases were acquired, apparently via HGT. The Lactobacillales ancestor was likely a microaerophile or an anaerobe, which is reflected in the loss of heme/copper-type cytochrome/quinol oxidase-related genes and catalase, characteristic enzymes of aerobic bacteria. In addition, several probable nonorthologous gene displacements via HGT were identified (Table 5). Furthermore, Lactobacillales (or the common ancestor of Bacilli) might have acquired the complete mevalonate pathway via HGT, possibly from an archaeal source (directly or through a bacterial intermediate). This pathway displaced the ancestral bacterial deoxyxylulose pathway of isoprenoid biosynthesis. The acquisition of the mevalonate pathway tree with a subsequent duplication of the mevalonate kinase gene is supported by the specific organization of the genes for four enzymes of this pathway [mevalonate and phosphomevalonate kinases, mevalonate pyrophosphate decarboxylase, and isopentenyl-diphosphate \(\delta\)-isomerase (LaCOGs 296, 298, 297, and 299, respectively)] in a single operon that is conserved in most Lactobacillales genomes.

In addition to the metabolic reduction, a major part of the gene loss in the common ancestor of Lactobacillales was sporulation-related functions encoded by the common ancestor of Bacilli. Despite the absence of genes for sporulation, catalase, and other key enzymes of oxidative stress response (e.g., superoxide dismutase) in 8 of the 12 genomes analyzed here (possibly multiple losses), at least some lactobacilli show enhanced stress resistance. This resistance is demonstrated by the increased recovery of live lactobacilli from vacuum-dried and irradiated food (30) by comparison to staphylococcal and Salmonella...
species. This resistance may be mediated in part by the low content of iron, a potent oxidant, which is accompanied by accumulation of manganese, a powerful antioxidant (31–33). Additional protection is likely to be provided by other antioxidants, including glutathione and γ-glutamylcysteine. Several Lactobacillus species encode a bifunctional glutathione synthetase (GshAB), whereas others have only γ-glutamylcysteine synthetase (GshA) (LaCOG01892). However, even lactococci that cannot synthesize glutathione have been shown to accumulate it, apparently via transport from the environment (34).

Loss of ancestral genes seems to be the prevailing trend in the evolution of Lactobacillales, as in other bacteria (35). Like all other bacterial lineages (36), lactobacilli also have a significant number of expanded gene families that evolved either by lineage-specific gene duplication or by acquisition of paralogous genes via HGT (37). A closer examination of these families indicates that adaptation to growth in nutrient-rich environments was the major driving force behind the fixation of duplications during the evolution of the Lactobacillales (Table 6, which is published as supporting information on the PNAS web site). An interesting case of ancient gene acquisition is the second enolase, which is characteristic of the Lactobacillales. All other bacteria have a single copy of this nearly ubiquitous glycolytic enzyme, but most of the Lactobacillales have two (with some differential gene loss). Phylogenetic analysis shows that one of these copies is the ancestral version in Gram-positive bacteria, whereas the other copy had been acquired by the ancestor of the Lactobacillales from a different bacterial lineage, most likely, Actinobacteria. The evolution rate of both enolases seems to be increased in the Lactobacillales branches of the phylogenetic tree (Fig. 6, which is published as supporting information on the PNAS web site). It has been shown that both enolases of Lc. lactis spp. lactis have enzymatic activity (38); however, their specific physiological functions remain unknown. Many other genes for proteins involved in sugar metabolism and transport were duplicated early in the evolution of the Lactobacillales, including phosphoenolpyruvate phosphotransferase systems, β-galactosidase, GpmB family sugar phosphatases, galactose mutarotase, and L-lactate dehydrogenases of two distinct classes. In addition to the apparent acquisition of new peptidases via HGT, duplications of several lineage-specific genes for these enzymes and for amino acid transporters were detected. Several paralogous expansions include genes for putative proteins related to those involved in antibiotic resistance in other bacteria, such as β-lactamases and penicillin V acylase. However, most LAB species analyzed here have been shown to be sensitive to common antibiotics and, after centuries of consumption by humans, have been, accordingly, “generally recognized as safe” (39, 40). Conceivably, the homologs of antibiotic-resistance genes are involved in normal cell-wall biosynthesis in the Lactobacillales. In the same context, expansion of a distinct family of tyrosine/serine phosphatases, which are often localized in the same operon with a serine/threonine protein kinase fused to several β-lactam-binding domains, is likely to be important for the regulation of cell-wall biosynthesis (41). Furthermore, Lactobacillales encode a paralog of class II lysyl-tRNA synthetase, which is fused to a membrane-associated domain (COG2898) implicated in oxacillin-like antibiotic resistance (42) and is probably involved in cell-wall biosynthesis.

The subsequent evolution of the Lactobacillales reveals ancestral gene loss and metabolic simplification but also a considerable number of lineage-specific duplications and acquisitions of unique genes. Numerous parallel gene losses, especially of genes coding for biosynthetic enzymes, were detected in the major branches of Lactobacillales, which presumably reflects similar environmental pressures. For instance, genes for serine and glycine biosynthesis were lost in the common ancestor of Lactobacillaceae and Leuconostocaceae; genes for biosynthesis of arginine and aromatic amino acids were lost independently in Lb. brevis, P. pentosaceus, O. oeni, and the Lb. casei–Lb. delbrueckii group; and several fatty-acid-biosynthesis genes were lost in the Lb. gasseri and Lb. johnsonii branch (Table 2). Lineage-specific gene loss was extensive in the evolution of all lineages of Lactobacillales, but several species were especially notable “losers.” In particular, S. thermophilus not only lost numerous genes but also exhibited many fresh pseudogenes, suggesting an active and ongoing process of genome decay, similarly reported for two different strains of the same species (5). Moreover, substantial gene loss (368 genes according to the present reconstruction) also occurred at the base of the streptococci–lactococci branch (including several genes involved in cell division that are conserved in most bacteria, such as crvB, mreB, mreC, and MinD). Other lineages particularly prone to gene loss are P. pentosaceus (487 genes lost) and the Leuconostoc and Oenococcus branch, with 381 genes lost at the base of the branch and considerable additional loss in each species. Substantial gene loss also occurred during the evolution of the Lb. delbrueckii group (Lb. delbrueckii, Lb. gasseri, and Lb. johnsonii), leading to additional genome reduction in Lb. gasseri and Lb. johnsonii (Fig. 3). In the species with larger genomes, such as Lb. plantarum and Lb. casei, the loss of ancestral genes was counterbalanced by the emergence of many new genes via duplication and HGT (Fig. 3).

Comparison of the number of genes lost or gained on a particular tree branch and the length of the corresponding branch reveals a pattern similar to that described previously for Proteobacteria (43). The number of gene losses (even when normalized by the size of the ancestral genome) strongly and significantly correlates with the branch length determined from sequence divergence (R = 0.68; P < 5 × 10−4), whereas the number of gene gains (again, regardless of normalization) does not show such a correlation (R = 0.10; P > 0.1). The clock-like behavior of gene loss is consistent with a large number of small-scale events, which are randomly distributed along the evolutionary path. This pattern suggests evolution under purifying selection. In contrast, the lack of such correlation for gene gain appears to involve relatively large batches of genes acquired at a time, with longer intervals between the acquisition events, perhaps because of positive selection.

In addition to the reconstruction of the ancestral gene sets, we compared the genome organizations of all of the sequenced Lactobacillales genomes with previously developed computational methods (44). Only closely related species showed significant genome colinearity above the level of individual operons, and there was virtually no large-scale conservation of gene order between the four major groups of the Lactobacillales (data not shown). Thus, the processes of gene loss and acquisition during the evolution of these bacteria were accompanied by extensive genome rearrangements.

**Phyletic Patterns and Central Metabolism Reconstruction.** Given the prominence of sugar metabolism and energy conversion systems in Lactobacillales, we examined the evolution of these systems through phyletic patterns, reflecting the presence or absence of genes in individual genomes in a manner similar to that described in ref. 45. Most of the genes involved in these functions are represented in all species (Fig. 7 and Table 7, which are published as supporting information on the PNAS web site). These genes include those coding for the downstream part of glycolysis, from glyceraldehyde-3P to pyruvate and pyruvate conversion to lactate and 2,3-butanediol. Lactate formation from acetyl-CoA; several reactions of the pentose–phosphate pathway; and the mannose-specific phosphotransferase system. Clearly, these enzymes are insufficient to completely define the metabolism of any individual species, and several reactions are specific to individual lineages. The presence/absence patterns of key enzymes involved in lactate fermentation poorly correlate with the phenotypes of the Lactobacillales (Table 7). However, it has been shown that under certain conditions Lactobacillales can switch between sole production of lactic acid and
the production of mixed end products, including acetic acid, lactic acid, ethanol, and CO₂ (46, 47).

The metabolic potential of the Lactobacillales is complemented by its predicted transport capabilities. In particular, amino acid uptake systems dominate over sugar and peptide uptake systems. Among the detected sugar uptake systems, those specific for oligosaccharides and glycosides outnumber those for free sugars. In addition, Lactobacillales encode a variety of predicted drug, peptide, and macromolecular efflux pumps, some of which are likely to be involved in intercellular signaling.

Other metabolic capabilities of Lactobacillales are listed in Table 8, which is published as supporting information on the PNAS web site. Generally, Lb. brevis, Lb. johnsonii, Lb. gasseri, Lb. delbrueckii, and P. pentosaceus have extremely narrow repertoires of biosynthetic pathways, whereas Lc. lactis ssp. lactis, Lc. lactis ssp. cremoris, Lb. plantarum, and Lc. mesenteroides retain a much broader biosynthetic repertoire.

The Bacteriocins. Lactobacillales are known for producing specific antimicrobial peptides, the bacteriocins (48, 49). Several proteins are responsible for the modification, export, and regulation of bacteriocin production and are often encoded in the same operon with the bacteriocins (48, 49). Because bacteriocins are small proteins with highly diverged sequences, they are often hard to identify by amino acid conservation. Therefore, genome context analysis is required for a more complete characterization of the bacteriocin repertoire. Among the Lactobacillales genomes analyzed here, seven have clustered genes for (putative) bacteriocins and associated proteins. Within these regions, we identified two prebacteriocin families. One family consists of precursors of a known bacteriocin, pediocin from P. pentosaceus, homologs of which also are present in Lc. mesenteroides and Lb. casei (LaCOG01709). The second family consists of previously unidentified putative bacteriocin precursors distantly related to Diversin V41 (50) and present in P. pentosaceus and Lb. johnsonii (LaCOG03352). In addition, numerous small ORFs located in the immediate vicinity of the putative genes for bacteriocins and associated proteins might encode novel peptides (Fig. 8, which is published as supporting information on the PNAS web site). Bacteriocin-production-related genes seem to be among those that are often transferred horizontally as indicated by the analysis of the respective phylogenetic trees and differences in the operon organization, even in closely related genomes.

Concluding Remarks. This work is an extensive comparative analysis of a compact group of relatively closely related prokaryotic genomes that show a gradient of sequence conservation. Loss of ancestral genes and metabolic simplification are the central trends of LAB evolution. Major gene loss already occurred at the stage of the common ancestor of Lactobacillales, which indicates early adaptation to nutritionally rich environments. However, genome degradation appears to be an ongoing process given that all species of Lactobacillales show loss of specific genes, and many possess numerous pseudogenes. Beyond gene loss, Lactobacillales have clear ancestral adaptations for nutritionally rich, microaerophilic environments, which include acquisition via HGT and duplication of genes for various enzymes and transporters of sugar and amino acid metabolism. The molecular systems responsible for the production of specific antimicrobials, such as the bacteriocins, are among other adaptations that become apparent through comparative genomic analysis, probably reflecting the long-term existence of Lactobacillales in complex microbial communities. Comparison of the genomes of Lactobacillales suggests that the milk-digesting phenotype evolved independently in different bacterial lineages. This phenotype apparently does not require a unique set of genes but rather emerged through assortment and adaptation of enzymes shared with other bacteria.

The comparative genomic analysis described here also suggests a revision of the taxonomy of the Lactobacillales. Phylogenetic analysis of multiple protein sequences showed that the streptococci–lactococci branch is basal in the Lactobacillales tree and that the Pediococcus group is a sister to the Leuconostoc group, which supports the paraphyly of the Lactobacillus genus. Furthermore, Lb. casei is confidently placed at the base of the Lb. delbrueckii group, which contradicts the earlier classification.

Materials and Methods

Whole-genome shotgun sequencing was carried out at the U.S. Department of Energy Joint Genome Institute. Genomes were sequenced to ~8 X coverage and assembled by using Jazz, the Joint Genome Institute assembler (51). Gap closure was carried out at Fidelity Systems, Inc., by using direct genomic sequencing (52).

ORFs were identified with the GeneMarkS program (53). Gene functions were predicted by assigning predicted genes to COGs (www.ncbi.nlm.nih.gov/COG) by using the COGNITOR method (24) and by database searches conducted with the PSI-BLAST program (54). Transfer RNAs were predicted with the tRNA-Scan-SE program (55). LaCOGs were constructed by using previously described procedures (23, 56). Phylogenetic analysis was performed by using the least-square or maximum-likelihood methods, and gene gain/loss scenarios were reconstructed with a version of the weighted parsimony algorithm (29).

Additional methodological details and a detailed list of data deposition numbers are provided in Supporting Materials and Methods, which is published as supporting information on the PNAS web site.

We dedicate this work to the career and many contributions of Prof. Larry McKay, a pioneering researcher on the genetics of lactic acid bacteria. This work was supported by the Intramural Program of the National Library of Medicine; the U.S. Department of Energy Office of Science, Biological, and Environmental Research Program and the University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48; the Lawrence Berkeley National Laboratory under Contract DE-AC03-76SF00098; and the Los Alamos National Laboratory under Contract W-7405-ENG-36. Genome closure was supported by Department of Energy Small Business Innovative Research Grants DE-FG02-98ER25277 and DE-FG02-00ER25309 (to A. Slesarev and S.K.). Additional funding came from the American Vineyard Foundation (to D.M.); the California Competitive Grants Program for Research in Viticulture and Enology (to D.M.); the Western Dairy Center (to B.W.); the Utah Agricultural Experiment Station (to B.W.); Dairy Management, Inc. (to B.W., T.K., and D.O.); Danisco, Inc. (to T.K.); the Southeast Dairy Foods Research Center (to T.K.); the North Carolina Dairy Foundation (to T.K.); the North Carolina Agricultural Research Service (to T.K.); National Institutes of Health Grant R01 GM55343 (to M.S.); the U.S. Department of Agriculture Agricultural Research Service Food Science Research Unit (to F.B.); Pickle Packers International, Inc. (to F.B.); the Utah State Community University Research Initiative (to J.B.); the Nebraska American Dairy Association (to R.H.); Minnesota Agriculture Experiment Station Project 18-062 (to K.B.); the Minnesota–South Dakota Dairy Foods Research Center (to D.O.); and the University of Wisconsin College of Agriculture and Life Science (to J.S.).


Figure 4S. Phylogenetic tree of Lactobacillales constructed on the basis of concatenated alignments of RNA polymerase subunits. All branches are supported at >75% bootstrap values. Species are colored according to the current taxonomy: Lactobacillaceae – blue; Leuconostocaceae – magenta; Streptococcaceae – red;
Figure 5S: Proportion of genes (56%) that evolve similarly (same mode) with ribosomal proteins.
Figure 6S. Phylogenetic analysis of Enolase. Maximum-likelihood unrooted tree was built using the MOLPHY program. The same program was used to compute bootstrap probabilities. Each terminal node of the tree is labeled by the numeric Genbank identifier (GI) number (where available) and the respective species name. Those major branches of interest that were supported by bootstrap probability greater than 70% are marked by black circles. The species analyzed in this work are shown in blue.
Figure 7S. A schematic representation of the reconstruction of keu metabolic pathways associated with central carbon (carbohydrate) metabolism in lactic acid bacteria.

Black arrows show reactions present in all species, pale blue - in all except one, red – in a smaller subset of species; for the latter category, phyletic patterns are indicated (the detailed information for all LaCOGs associated with this figure is provided in the Table 2B, supporting information).

In the phyletic patterns, "|" indicates presence of the gene in a given species and "-" indicates absence. Species in the phyletic pattern are shown in the order: Streptococcus thermophilus, Lactococcus lactis ssp. lactis, Lactococcus lactis ssp. cremoris, Lactobacillus brevis, Lactobacillus plantarum, Pediococcus pentosaceus, Leuconostoc mesenteroides, Oenococcus oeni, Lactobacillus johnsonii, Lactobacillus casei. The systematic protein names (from E. coli, B. subtilis) or Lactobacillales for the enzymes assigned to each reaction are indicated. Key reactions of homo- and heterofermentation are color-coded (green and pink, respectively).

Substrates that are additional precursors or product of several reactions are dark green.

+ - Phyletic pattern for phosphoenolpyruvate carboxykinase, pckA, includes representatives of LaCOG2238 and a single protein from L. casei;

* - Acetoin reductase, ButA, homologs belong to the large family of short chain dehydrogenases, from multiple LaCOGs. Many of them have unknown substrate specificity and might be involved in the same reaction as ButA;

& - While there is no dedicated PTS system for lactose transport identified, it is possible that some PTS systems with wide substrate specificity also can transport lactose.

^ - Presence the system was essentially determined on the basis of the presence of this gene.
**Figure 8S** Genome clusters encoding bacteriocins and genes for their export systems, including novel putative bacteriocins