Comparative genomics of the lactic acid bacteria

K. Makarova[a], A. Slesarev[b], Y. Wolf[a], A. Sorokin[a], B. Mirkin[c], E. Koonin[a,d], A. Pavlov[b], N. Pavlova[b], V. Karamychev[b], N. Polouchine[b], V. Shakhova[b], I. Grigoriev[e], Y. Lou[e], D. Rohksar[e], S. Lucas[e], K. Huang[e,f], D. M. Goodstein[e], T. Hawkins[e,f], V. Plengvidhya[f,g,h], D. Welker[i], J. Hughes[i], Y. Goh[j], A. Benson[j], K. Baldwin[k], J.-H. Lee[k], I. Díaz-Muñiz[f,l], B. Dosti[l], V. Smeianov[l], W. Wechter[f,l], R. Barabote[m], G. Lorca[f,m], E. Altermann[f,g], R. Barrangou[f,g], B. Ganesan[n,o], Y. Xie[f,n,o], H. Rawsthorne[f,p], D. Tamir[f,p], C. Parker[f,p], F. Breidt[g,h], J. Broadbent[o], R. Hutkins[j], D. O'Sullivan[k], J. Steele[l], G. Unlu[q], M. Saier[m], T. Klaenhammer[d,g], P. Richardson[e], S. Kozyavkin[b], B. Weimer[d,n,o], and D. Mills[d,p]

[a]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894; [b]Fidelity Systems Inc., 7961 Cessna Avenue, Gaithersburg, MD 20879; [c]School of Information Systems and Computer Science, Birkbeck College, University of London, Malet Street, London WC1E 7HX, United Kingdom; [e]U.S. Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598; [g]Department of Food Science, North Carolina State University, Raleigh, NC 27695; [h]North Carolina Agricultural Research Service, U.S. Department of Agriculture, Raleigh, NC 27695; Departments of [i]Biology and [o]Nutrition and Food Science and [n]Center for Integrated BioSystems, Utah State University, Logan, UT 84322; [j]Department of Food Science and Technology, University of Nebraska, Lincoln, NE 68583; [k]Department of Food Science and Nutrition, University of Minnesota, St. Paul, MN 55108; [l]Department of Food Science, University of Wisconsin, Madison, WI 53706; [m]Department of Biology, University of California at San Diego, La Jolla, CA 92093; [p]Department of Viticulture and Enology, University of California, Davis, CA 95616; and [q]Department of Food Science and Toxicology, University of Idaho, Moscow, ID 83844

Contributed by T. Klaenhammer, August 16, 2006

Lactic acid-producing bacteria 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.

evolutionary genomics | fermentation

Lactic acid bacteria (LAB) are historically defined as a group of microaerophilic, Gram-positive organisms that ferment hexose sugars to produce primarily lactic acid. This functional classification includes a variety of industrially important genera, including *Lactococcus*, *Enterococcus*, *Oenococcus*, *Pediococcus*, *Streptococcus*, *Leuconostoc*, and *Lactobacillus* species. The seemingly simplistic metabolism of LAB has been exploited throughout history for the preservation of foods and beverages in nearly all societies dating back to the origins of agriculture (1). Domestication of LAB strains passed down through various culinary traditions and continuous passage on food stuffs has resulted in modern-day cultures able to carry out these fermentations. Today, LAB play a prominent role in the world food supply, performing the main bioconversions in fermented dairy products, meats, and vegetables. LAB also are critical for the production of wine, coffee, silage, cocoa, sourdough, and numerous indigenous food fermentations (2).

LAB species are indigenous to food-related habitats, including plant (fruits, vegetables, and cereal grains) and milk environments. In addition, LAB are naturally associated with the mucosal surfaces of animals, e.g., small intestine, colon, and vagina. Isolates of the same species often are obtained from plant, dairy, and animal habitats, implying wide distribution and specialized adaptation to these diverse environments. LAB species employ two pathways to metabolize hexose: a homofermentative pathway in which lactic acid is the primary product and a heterofermentative pathway in which lactic acid, CO_2, acetic acid, and/or ethanol are produced (3).

Complete genome sequences have been published for eight fermentative and commensal LAB species: *Lactococcus lactis*, *Lactobacillus plantarum*, *Lactobacillus johnsonii*, *Lactobacillus acidophilus*, *Lactobacillus sakei*, *Lactobacillus bulgaricus*, *Lactobacillus salivarius*, and *Streptococcus thermophilus* (4–11). This study examines nine other LAB genomes representing the phylogenetic and functional diversity of lactic acid-producing microorganisms. The LAB have small genomes encoding a range of biosynthetic capabilities that reflect both prototrophic and auxotrophic characters. Phylogenetic analyses, comparison of genomic content across the group, and reconstruction of ancestral gene sets reveal a combination of gene loss and gain during the coevolution of LAB with animals and the foods they consumed.

Results and Discussion

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 the PNAS 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* (5) 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). All

Author contributions: A. Slesarev, E.K., T.H., F.B., J.B., R.H., D.O., J.S., G.U., M.S., T.K., P.R., S.K., B.W., and D.M. designed research; K.M., A. Slesarev, A. Sorokin, B.M., A.P., N. Pavlova, V.K., N. Polouchine, V. Shakhova, I.G., Y.L., S.L., K.H., D.M.G., V.P., D.W., J.H., Y.G., K.B., J.-H.L., I.D.-M., B.D., V. Smeianov, W.W., R. Barabote, G.L., E.A., R. Barrangou, B.G., Y.X., H.R., D.T., C.P., and S.K. performed research; A.P., N. Pavlova, N. Polouchine, B.G., Y.X., and S.K. contributed new reagents/analytic tools; A. Slesarev, Y.W., A.P., V.K., I.G., D.R., A.B., G.L., E.A., R. Barrangou, F.B., J.B., R.H., D.O., J.S., G.U., M.S., T.K., P.R., S.K., B.W., and D.M. analyzed data; and K.M., Y.W., E.K., T.K., P.R., S.K., and D.M. wrote the paper.

The authors declare no conflict of interest.

Abbreviations: LAB, lactic acid bacteria; COG, cluster of orthologous genes; LaCOG, *Lactobacillales*-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).

[d]To whom correspondence may be addressed. E-mail: koonin@ncbi.,nlm.nih.gov, trk@unity.ncsu.edu, bcweimer@cc.usu.edu, or damills@ucdavis.edu.

[f]Present addresses for K.H., T.H., V.P., I.D.-M., W.W., G.L., E.A., R. Barrangou, Y.X., H.R., D.T., and C.P. are published as supporting information on the PNAS web site.

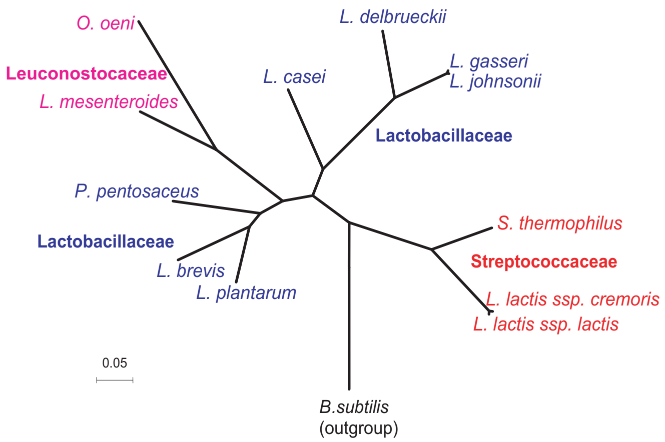

Fig. 1. Phylogenetic trees of *Lactobacillales* constructed on the basis of concatenated alignments of ribosomal proteins. All branches are supported at >75% bootstrap values. Species are colored according to the current taxonomy: *Lactobacillaceae*, blue; *Leuconostocaceae*, magenta; *Streptococcaceae*,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

LAB genomes contained transposons, ranging from ≈0.2% of the genome in *Lactobacillus gasseri* to nearly 5% in *Lc. 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 (*Le. 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 *Lc. lactis* ssp. *cremoris*) formed a separate branch (16).

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 nonsynonymous 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 nonsynonymous sites) ratio was 38.5 ± 0.5 for the *Lb. gasseri*/*Lb. johnsonii* pair and 29.8 ± 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-thiouridylate) methyltransferase and

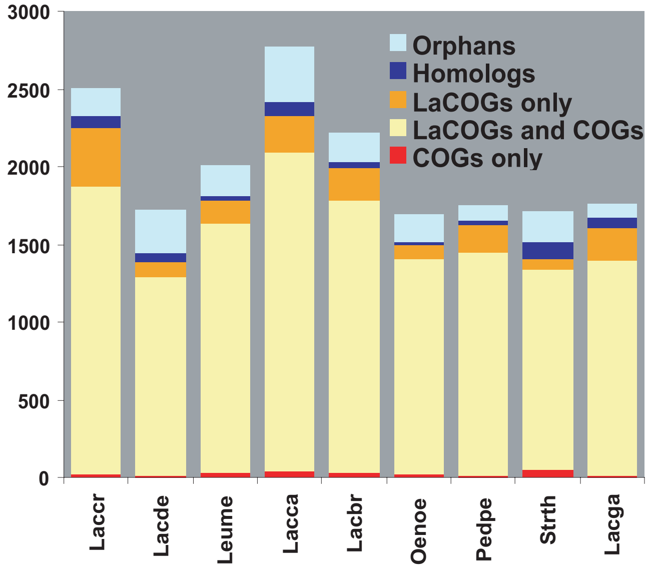

Fig. 2. Conserved and unique genes in the genomes of *Lactobacillales*. "Homolog" indicates genes with detectable homologs in organisms other than those analyzed here but could not be included in any of the COG sets. "Orphans" are predicted genes without detectable homologs. Lacga, *Lb. gasseri*; Lacbr, *Lb. brevis*; Pedpe, *P. pentosaceus*; Laccr, *Lc. lactis* ssp. *cremoris*; Strth, *S. thermophilus*; Oenoe, *O. oeni*; Leume, *Le. mesenteroides*; Lacca, *Lb. casei*; Lacde, *Lb. delbrueckii*; Lacla, *Lc. lactis*; Lacpl, *Lb. plantarum*; Lacjo, *Lb. johnsonii*. The vertical axis shows the number of genes per genome.

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 evolution of ribosomal proteins, a linear dependence between the two matrices is observed. At least ≈25% of the LaCOGs showed strong deviations from the linear dependence, suggesting a high level of HGT and/or major local accelerations of evolution (Fig. 5, 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). Of particular interest is the substantially elevated level of such violations among genes involved in sugar metabolism, including key enzymes, such as phosphoketolase, transketolase, and various components of phosphotransferase systems (Table 4, which is published as supporting information on the PNAS web site). HGT via bacteriophage-mediated or conjugative mechanisms has been extensively documented in the *Lactobacillales* and is a prominent process for niche-specific adaptation in the lactococci (27). Noteworthy is the potential for such transfer in *Lc. lactis* ssp. *cremoris* 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 ≈2,100–2,200 genes, losing 600–1,200 genes (≈25–30%) and gaining <100 genes compared with the ancestor of all *Bacilli*, for which the genome size of ≈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 δ-isomerase (LaCOGs 296, 298, 297, and 299, respectively)] in a single operon that is conserved in most *Lactobacillales* genomes.

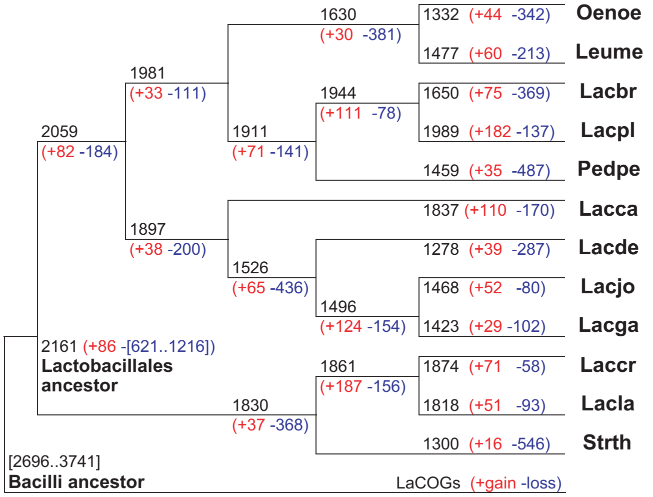

Fig. 3. Reconstruction of gene content evolution in *Lactobacillales*. The tree topology is as in Fig. 1, rooted by using *Baccillus subtilis* as the outgroup. For each species and each internal node of tree, the inferred number of LaCOGs present, and the numbers of LaCOGs lost (blue) and gained (red) along the branch leading to the given node (species) are indicated. Abbreviations are as in Fig. 2.

In addition to the metabolic reduction, a major part of the gene loss in the common ancestor of *Lactobacillales* were 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* ssp. *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 *crcB*, *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 ($R = 0.68$; $P < 5 \times 10^{-4}$), whereas the number of gene gains (again, regardless of normalization) does not show such a correlation ($R = 0.16$; $P > 0.1$). 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-butandiol; acetate 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 CO_2 (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 *Le. 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 *Le. mesenteroides* and *Lb. casei* (LaCOG01709). The second family consists of previously unidentified putative bacteriocin precursors distantly related to Divercin 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× depth 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 tRNAscan-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-AC03-76SF00098l; 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-FG02-98ER82577 and DE-FG02-00ER83009 (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 GM55434 (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 Agriculuture and Life Science (to J.S.).

1. Miller N, Wetterstrom W (2000) in *The Cambridge World History of Food*, eds Kiple K, Ornelas K (Cambridge Univ Press, Cambridge, UK), Vol 2, pp 1123–1139.
2. Wood B (1998) *Microbiology of Fermented Foods* (Blackie, London).
3. Kandler O (1983) *Antonie van Leeuwenhoek* 49:209–224.
4. Altermann E, Russell W, Azcarate-Peril M, Barrangou R, Buck B, McAuliffe O, Souther N, Dobson A, Duong T, Callanan M, *et al*. (2005) *Proc Natl Acad Sci USA* 102:3906–3912.
5. Bolotin A, Quinquis B, Renault P, Sorokin A, Ehrlich S, Kulakauskas S, Lapidus A, Goltsman E, Mazur M, Pusch G, *et al*. (2004) *Nat Biotechnol* 22:1554–1558.

6. Bolotin A, Wincker P, Mauger S, Jaillon O, Malarme K, Weissenbach J, Ehrlich S, Sorokin A (2001) *Genome Res* 11:731–753.
7. Chaillou S, Champomier-Verges M, Cornet M, Crutz-Le Coq A, Dudez AM, Martin V, Beaufils S, Darbon-Rongere E, Bossy R, Loux V, Zagorec M (2005) *Nat Biotechnol* 23:1527–1533.
8. Kleerebezem M, Boekhorst J, van Kranenburg R, Molenaar D, Kuipers OP, Leer R, Tarchini R, Peters S, Sandbrink H, Fiers M, *et al*. (2003) *Proc Natl Acad Sci USA* 100:1990–1995.
9. Pridmore R, Berger B, Desiere F, Vilanova D, Barretto C, Pittet A, Zwahlen M, Rouvet M, Altermann E, Barrangou R, *et al*. (2004) *Proc Natl Acad Sci USA* 101:2512–2517.
10. Claesson M, Li Y, Leahy S, Canchaya C, van Pijkeren J, Cerdeno-Tarraga A, Parkhill J, Flynn S, O'Sullivan G, Collins J, *et al*. (2006) *Proc Natl Acad Sci USA* 103:6718–6723.
11. van de Guchte M, Penaud S, Grimaldi C, Barbe V, Bryson K, Nicolas P, Robert C, Oztas S, Mangenot S, Couloux A, *et al*. (2006) *Proc Natl Acad Sci USA* 103:9274–9279.
12. Klappenbach J, Dunbar J, Schmidt T (2000) *Appl Environ Microbiol* 66:1328–1333.
13. Di Mattia E, Grego S, Cacciari I (2002) *Microb Ecol* 43:34–43.
14. McKay L, Baldwin K (1990) *FEMS Microbiol Rev* 7:3–14.
15. Vandamme P, Pot B, Gillis M, de Vos P, Kersters K, Swings J (1996) *Microbiol Rev* 60:407–438.
16. Wood B, Holzapfel W (1995) *The Genera of Lactic Acid Bacteria* (Blackie, Glasgow, UK).
17. Wolf Y, Rogozin I, Grishin N, Tatusov R, Koonin E (2001) *BMC Evol Biol* 1:8.
18. Siezen R, van Enckevort F, Kleerebezem M, Teusink B (2004) *Curr Opin Biotechnol* 15:105–115.
19. Takezaki N, Rzhetsky A, Nei M (1995) *Mol Biol Evol* 12:823–833.
20. Yang D, Woese C (1989) *Syst Appl Microbiol* 12:145–149.
21. Jordan I, Rogozin I, Wolf Y, Koonin E (2002) *Theor Popul Biol* 61:435–447.
22. Lynch M, Conery J (2003) *Science* 302:1401–1404.
23. Tatusov R, Koonin E, Lipman D (1997) *Science* 278:631–637.
24. Tatusov R, Fedorova N, Jackson J, Jacobs A, Kiryutin B, Koonin E, Krylov D, Mazumder R, Mekhedov S, Nikolskaya A, *et al*. (2003) *BMC Bioinformatics* 4:41.
25. Leipuviene R, Qian Q, Bjork G (2004) *J Bacteriol* 186:758–766.
26. Novichkov P, Omelchenko M, Gelfand M, Mironov A, Wolf Y, Koonin E (2004) *J Bacteriol* 186:6575–6585.
27. Wood B, Warner P (2003) *Genetics of Lactic Acid Bacteria* (Kluwer Academic/Plenum, New York).
28. Siezen R, Renckens B, van Swam I, Peters S, van Kranenburg R, Kleerebezem M, de Vos W (2005) *Appl Environ Microbiol* 71:8371–8382.
29. Mirkin B, Fenner T, Galperin M, Koonin E (2003) *BMC Evol Biol* 3:2.
30. Hastings J, Holzapfel W, Niemand J (1986) *Appl Environ Microbiol* 52:898–901.
31. Horsburgh M, Wharton S, Karavolos M, Foster S (2002) *Trends Microbiol* 10:496–501.
32. Daly M, Gaidamakova E, Matrosova V, Vasilenko A, Zhai M, Venkateswaran A, Hess M, Omelchenko M, Kostandarithes H, Makarova KS, *et al*. (2004) *Science* 306:1025–1028.
33. Archibald F, Fridovich I (1981) *J Bacteriol* 145:442–451.
34. Li Y, Hugenholtz J, Abee T, Molenaar D (2003) *Appl Environ Microbiol* 69:5739–5745.
35. Ochman H (2005) *Proc Natl Acad Sci USA* 102:11959–11960.
36. Jordan I, Makarova K, Spouge J, Wolf Y, Koonin E (2001) *Genome Res* 11:555–565.
37. Koonin E, Makarova K, Aravind L (2001) *Annu Rev Microbiol 55:709–742.*
38. Jamet E, Ehrlich S, Duperray F, Renault P (2001) *Lait* 81:115–129.
39. Katla A, Kruse H, Johnsen G, Herikstad H (2001) *Int J Food Microbiol* 67:147–152.
40. Teuber M, Meile L, Schwarz F (1999) *Antonie van Leeuwenhoek* 76:115–137.
41. Yeats C, Finn RD, Bateman A (2002) *Trends Biochem Sci* 27:438.
42. Nishi H, Komatsuzawa H, Fujiwara T, McCallum N, Sugai M (2004) *Antimicrob Agents Chemother* 48:4800–4807.
43. Snel B, Bork P, Huynen M (2002) *Genome Res* 12:17–25.
44. Wolf Y, Rogozin I, Kondrashov A, Koonin E (2001) *Genome Res* 11:356–372.
45. Koonin E, Galperin M (2003) *Sequence–Evolution–Function: Computational Approaches in Comparative Genomics* (Kluwer Academic, London).
46. Hemme D, Foucaud-Scheunemann C (2004) *Int Dairy J* 14:467–494.
47. Liu S-Q (2003) *Int J Food Microbiol* 83:115–131.
48. Twomey D, Ross R, Ryan M, Meaney B, Hill C (2002) *Antonie van Leeuwenhoek* 82:165–185.
49. Nes I, Johnsborg O (2004) *Curr Opin Biotechnol* 15:100–104.
50. Metivier A, Pilet M, Dousset X, Sorokine O, Anglade P, Zagorec M, Piard JC, Marion D, Cenatiempo Y, Fremaux C (1998) *Microbiology* 144:2837–2844.
51. Aparicio S, Chapman J, Stupka E, Putnam N, Chia J, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, *et al*. (2002) *Science* 297:1301–1310.
52. Slesarev A, Mezhevaya K, Makarova K, Polushin N, Shcherbinina O, Shakhova V, Belova G, Aravind L, Natale D, Rogozin I, *et al*. (2002) *Proc Natl Acad Sci USA* 99:4644–4649.
53. Besemer J, Lomsadze A, Borodovsky M (2001) *Nucleic Acids Res* 29:2607–2618.
54. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D (1997) *Nucleic Acids Res* 25:3389–3402.
55. Lowe T, Eddy S (1997) *Nucleic Acids Res* 25:955–964.
56. Tatusov R, Natale D, Garkavtsev I, Tatusova T, Shankavaram U, Rao BS, Kiryutin B, Galperin M, Fedorova N, Koonin E (2001) *Nucleic Acids Res* 29:22–28.