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. Loue, 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. Dı´az-Muniz[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. Ganesann,[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. Weimerd,[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
[q]Department of Food Science and Toxicology, University of Idaho, Moscow, ID 83844

October 2006

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

DISCLAIMER

Comparative genomics of the lactic acid bacteria

D. A. Mills[1†], K. Makarova[2], B. Weimer[3†], Y. Wolf[2], E. Koonin[2], T. R. Klaenhammer[4†], M. Saier[5†], A. Sorokin[2], B. Mirkin[x], I. Grigoriev[6], Y. Lou[6], P. Richardson[6], D. Rohksar[6], S. M. Lucas[6], K. Huang[6μ], D. M. Goodstein[6], T. H. Hawkins[6¥], F. Breidt[4,7†], J. Broadbent[8†], R. Hutkins[9†], L. McKay[10†], D. O'Sullivan[10, 11†], J. Steele[12†], G. Yuksel[13†], H. Rawsthorne[1§], D. Tamir[1£], C. Parker[1‡], X. Yie[3∫], B. Ganesan[3], E. Alterman[4], R. Barrangou[4], G. Lorca[5π], R. Barabote[5], V. Plengvidhya[4,7¢], D. Welker[14], J. Hughes[14], Y. Goh[9], A. Benson[9], K. Baldwin[10], J.-H. Lee[10], I. Dias[12∂], B. Dosti[12], V. Smeianov[12], W. Wechter[12∞], A. Slesarev[15], S. Kozyavkin[15].

1. Department of Viticulture and Enology, University of California, Davis, CA 95616.
2. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894.
3. Center for Integrated BioSystems, Center for Microbe Detection & Physiology, Utah State University, Logon, UT 84322
4. Department of Food Science, Southeast Dairy Foods Research Center, North Carolina State University, Raleigh, NC. 27695
5. Department of Biology, University of California, San Diego, CA.
6. U.S. DoE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598.
7. U.S. Department of Agriculture, Agricultural Research Service, North Carolina Agricultural Research Service, Raleigh, NC 27695
8. Department of Nutrition and Food Science, Utah State University, Logon, UT 84322.
9. Department of Food Science and Technology, University of Nebraska-Lincoln, Lincoln, NE 68583.
10. Department of Food Science and Nutrition, University of Minnesota, St. Paul, MN 55108.
11. Center for Microbial and Plant Genomics, University of Minnesota, St. Paul, MN 55108.
12. Department of Food Science, University of Wisconsin-Madison, Madison, WI 53706.
13. Department of Food Science and Toxicology, University of Idaho, Moscow, ID 83844.
14. Department of Biology, Utah State University, Logon, UT 84322.
15. Fidelity Systems Inc., 7961 Cessna Avenue, Gaithersburg, MD 20879.

x School of Information Systems and Computer Science, Birkbeck College, University of London, Malet Street, London, WC1E 7HX, UK

[†]Lactic Acid Bacteria Genomics Consortium

[μ]Present address: Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley CA 94720.

[¥]Present address: GE Healthcare, 800 Centennial Avenue, Piscataway, NJ 08855

[§]Present address: Dept. Bioscience and Biotechnology, Drexel University, 32nd & Chestnut St. Philadelphia, PA 19104-2875.

[£]Present address: Department of Plant Pathology and Microbiology, Agricultural, Food and Environmental Quality Sciences, The Hebrew University of Jerusalem, P.O.Box 12, Rehovot 76100, Israel
[‡]Present address: Osel, Inc., 1800 Wyatt Dr. Suite 14, Santa Clara, CA 95054
[ʃ]Present address: Johns Hopkins University, Department of Pediatric Infectious Diseases, 720 Rutland Avenue, Ross 1109, Baltimore, Maryland 21205
[π]Present address: Banting and Best Institute, University of Toronto, 112 College St. M5G 1L6 Toronto, Canada. M5G 1L6
[¢]Present address: National Center for Genetic Engineering and Biotechnology (BIOTEC), 113 Paholyothin Rd., Klong 1, Klong Luang, Pathumthani 12120, Thailand
[∂]Present address: U.S. Department of Agriculture, Agricultural Research Service, North Carolina Agricultural Research Service, Department of Food Science, NC State University, Raleigh, NC 27695
[∞]Present address: US Vegetable Laboratory, USDA, ARS, Charleston, SC 29414

Introduction

The lactic acid bacteria (LAB) are historically defined as a group of microaerophillic, 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 a variety of *Lactobacillus* species. Bifidobacterium, a phylogenetically distinct genus that also produces lactic acid, has often been historically grouped with the LAB. The LAB have been employed throughout history for production of a range of fermented foods, performing the main bioconversion in fermented dairy products, meats and vegetables as well as critical bioconversions in the production of wine, coffee, silage, cocoa, and sourdough. These fermentations play a prominent role in the world food supply.

Various LAB species are indigenous to food-related habitats including plant (fruits, vegetables, cereal grains) and milk environments. In addition LAB are commensals that naturally associate with mucosal surfaces of animals, occupying varying locations (e.g. small intestine, colon, vagina). Often isolates of the same species may be obtained from plant, dairy and animal habitats, implying a wide distribution and adaptation to these differing environments. LAB species employ two different pathways to metabolize hexoses; a homofermentive pathway in which lactic acid is the primary product, a heterofermentative pathway in which equimolar amounts of lactic acid, acetic acid (or ethanol) are produced. Bifidobacteria employ a different pathway, "the bifidus pathway" which produces acetic and lactic acids. The LAB group includes both thermotolerant and mesophilic species as well as members that are highly resistant to acid and ethanol, promoting their activity and survival in acidified foods and alcoholic beverages.

To date, complete genome sequences have been published for five LAB species, representing both fermentative and commensal species; *Lactococcus lactis, Lactobacillus plantarum, Lactobacillus johnsonii, Lactobacillus acidophilus,* and *Streptococcus thermophilus* {Bolotin, 2004 #407; Bolotin, 2001 #409; Kleerebezem, 2003 #408; Pridmore, 2004 #395;Altermann, 2005 #440} as well as a *Bifidobacterium longum* species {Schell, 2002 #441}. This study examines nine new LAB genomes and a new *B. longum* genome 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 pathways. Collectively, the LAB display extensive transporter capacities for efficient carbon and nitrogen acquisition from the nutritionally rich environments they commonly inhabit. Phylogenetic analyses, comparison of genomic content across the group and reconstruction of ancestral gene sets indicate a combination of gene loss and key gene acquisitions during co-evolution of LAB with animals and the foods they consumed.

Results and discussion

General features of the LAB genomes

The gross features of the sequenced LAB genomes are summarized in Table 1. The number of predicted protein-coding genes in the LAB differs from ~1700 to ~2800. Given the close phylogenetic affinity of these organisms, such a difference suggests substantial gene loss and/or gain in their evolution (see below). In addition, all LAB genomes harbor pseudogenes. Strikingly, the number of pseudogenes differs by an order of magnitude, from <20 in *L. mesenteroides* and *P. pentosaceus* to ~200 in S. thermophilus and *L. delbrueckii*, suggesting an active, ongoing process of genome degeneration. The LABs also differ in the number of rRNA genes, from 2 in *O. oeni* to 9 *L. delbrueckii*, which correlates with the number of tRNA genes (Table 1) and might reflect differences in the ecological strategy (e.g., capacity for rapid growth) between these bacteria {Klappenbach, 2000 #442; Di Mattia, 2002 #443}. All LAB genomes contain transposons the content of which differs from ~0.2% of the genome in *L. gasseri* to nearly 5% in *L. lactis cremoris*. The underlying reasons behind such dramatic differences remain unknown and, again, might have to do with the distinct population structures and growth characteristics of different LAB. The majority of the LAB harbor plasmids of widely different sizes some of which are essential for the bacterial growth and carry genes for metabolic pathways, membrane transport and antibiotic resistance complementing the respective chromosomal genes ({McKay, 1990 #444}; Table 1).

Phylogenetic analysis and assessment of horizontal gene transfer impact on LAB evolution

Most of the LAB (Lacbr, Lacca, Laccr, Lacde, Lacga, Lacjo, Lacla, Lacpl, Leume, Oenoe, Pedpe and Strte) analyzed here belong to Firmicutes/Bacilli/Lactobacillales. The taxonomy of Lactobacillales remains an unresolved issue, in particular, because phenotypic classification (based on the type of fermentation) does not match the phylogeny based on DNA/RNA similarity {Vandamme, 1996 #391}. The rRNA tree of this group is poorly resolved {Vandamme, 1996 #391}, probably, because of high heterogeneity of evolutionary rates. Whole genome DNA and DNA-RNA hybridization and GC content studies led to delineation of three closely related lineages within Lactobacillales: the Leuconostoc group. (Leume, Oenoe), the *L. casei*-Pediococcus group (Lacpl, Lacca, Pedpe, Lacbr), and the *L. delbruekii* group (Lacde, Lacga, Lacjo) {Vandamme, 1996 #391}; streptococci (Strth) and lactococci (Lacla, Laccr) form a separate branch {, 1995 #400}.

The availability of complete genomes for all major branches of Lactobacillales enables a more definitive analysis of their evolutionary relationships. For this purpose, we constructed phylogenetic trees from concatenated protein sequences, an approach shown to improve the resolution and increase robustness of phylogenetic analysis {Wolf, 2001 #368}. Concatenated ribosomal proteins are considered to be the most reliable source for this type of phylogenetic analysis because, with a few exceptions, they are not prone to frequent horizontal gene transfer {Wolf, 2001 #368}; we supplemented the ribosomal protein dataset with concatenated RNA polymerase subunits which also undergo little if any horizontal transfer. Both trees, reconstructed with a variety of methods, display the same topology with strongly supported internal branches (**Fig.1**).

The Streptococci-Lactococci branch is basal in the Lactobacillales tree, and the Pediococcus group is a sister to the Leuconostoc group, which supports paraphyly of the Lactobacillus genus. Interestingly, *L. casei* is confidently placed at the base of the *L. delbruekii* group, which contradicts the earlier classification.

A molecular clock test {Takezaki, 1995 #418} showed a high heterogeneity of evolutionary rates within Lactobacillales (suppl. Figure **1S**). Most of the root-to-tip distances are significantly unequal to the mean tree height; the previously reported {Yang, 1989 #411} accelerated evolution of the Leuconostoc group (by a factor of 1.7-1.9 relative to the sister Pediococcus group) was especially prominent.

Whole-genome similarity measures – median distance between orthologs, gene content similarity and conservation of local gene context – suggest somewhat different relationships within Lactobacillales, with the Leuconostoc group being a sister to the Lactobacillus-Pediococcus lineages and, at least in some cases, *L. casei* returning to the *L. casei*-Pediococcus group in a manner consistent with the earlier classification {Vandamme, 1996 #391} (suppl. Figures**. 2S, 3S**). However, the latter two approaches to evolutionary reconstruction are not strictly phylogenetic and are strongly affected by horizontal gene transfer, parallel gene loss, and heterogeneity of evolutionary rates {Wolf, 2001 #368}. The basal position of the rapidly evolving Leuconostoc group relative to the Lactobacillus-Pediococcus group seen in the trees obtained with these methods is especially suggestive of a long-branch attraction artifact, given the aforementioned rapid evolution of Leuconostocacea. Conceivably, rapid sequence evolution in this group could be accompanied also by extensive gene loss and genome rearrangement (see below).

The level of purifying selection (in other words, the strength of selective constraints) acting on Lactobacillales species, can be estimated using two closely related pairs of genomes: *L. gasseri* - *L. johnsonii* and *L. lactis* subsp. *lactis* - *L. lactis* subsp. *cremoris*. Synonymous and non-synonymous substitution rates were estimated from concatenated coding sequence alignments of 443 LaCOGs which have only one ortholog from each of the four genomes (142031 codons). The d_S/d_N ratio was 37.9 for the *L. gasseri* - *L. johnsonii* pair and 28.2 for *L. l. lactis* - *L. l. cremoris* pair, showing unusually strong selective pressure as compared, e.g., to Proteobacteria which have characteristic d_S/d_N of 5-10 {Jordan, 2002 #438}. This potentially might reflect high effective population size of the LAB which is linked with the intensity of purifying selection {Lynch, 2003 #439}.

Reconstruction of clusters of orthologous groups in Lactobacillales

Robust identification of sets of orthologs (genes derived from the same ancestral gene) is a pre-requisite for informative evolutionary-genomic analysis of any group of organisms. We constructed a set of Lactobacillales-specific clusters of orthologous genes (COGs or LaCOGs for the Lactobacillales-specific set) from proteins encoded in the 12 sequenced genomes[xx] using the previously described computational procedure based on the evolutionary parsimony principle{Mirkin, 2003 #200}. Altogether, 3204 LaCOGs, which included from two to all 12 species, were identified. On average, LaCOGs covered 86% of the genome; 711 (22%) of LaCOGs could not be assigned to any COG previously identified in other organisms {Tatusov, 2001 #116} and therefore represented the fraction

of genes which, at this stage of genome sequencing, appear to be specific for Lactobacillales (**Figure 2**).

The fully conserved core of the analyzed Lactobacillales genomes, i.e., genes present in all 12 species (suppl. table "CORE"), consisted of 567 LaCOGs (18%). Examination of the distribution of the LaCOGs by function in the conserved core shows that the majority of these genes, as expected, encode components of the information-processing systems (translation, transcription, and replication; suppl. Fig and table "CORE"). However, the core set also includes a considerable number of uncharacterized genes (41) and genes with only a general prediction of biochemical activity (50). Since these genes are conserved throughout the Lactobacillales, it seems likely that they have essential functions, at least in this group, emphasizing the incompleteness of the current understanding of even the central cellular functions of relatively simple bacteria. Furthermore, for three core genes, no orthologs are detectable outside lactobacilli. One of these, ComX (LaCOG01447), is distantly related to Region 2 of sigma-70 and, indeed, is a competence regulator {Magnuson, 1994 #392}, another one contains a LysM (peptidoglycan-binding) domain (LaCOG01826), the exact function of which is unknown, and the third one (LaCOG01237) is so far unique to Lactobacillaceae and does not have any recognizable sequence features or genome context associations.

Local molecular clock and horizontal gene transfer

We tested the consistency of local molecular clock in individual LaCOGs using a recently developed technique {Novichkov, 2004 #402}. A matrix of inter-species distances for a given LaCOG was compared to the matrix of baseline distances obtained from the concatenated alignment of ribosomal proteins. If a COG evolves in a clock-like manner, a linear dependence between the two matrices is observed (i.e., when the two sets of distances are plotted against each other, all the points fit a straight line). At least ~25% of the LaCOGs showed strong deviation from the baseline, which suggests a high level of HGT and/or major, local accelerations of evolution (**Fig. 1S**). Several functional groups of genes show statistically significant differences in their propensity to local molecular clock violation (suppl. "**Clock and functional groups");** of particular interest is the substantially elevated level of apparent HGT among genes involved in sugar metabolism, including such key enzymes as components of various PTS, phosphoketolase, transketolase and others. (suppl. "**Sugar metabolism and HGT**").

Reconstruction of gene gain and loss events in the evolution of Lactobacillales.

We used a version of the weighted parsimony algorithm {Mirkin, 2003 #200} to reconstruct the events that occurred during the evolution of this group after its divergence from the common ancestor of all Bacilli (see Methods). The results are schematically shown in **Fig. 3**. This reconstruction suggests that the common ancestor of Lactobacillales had ~2100-2200 genes, losing 600-1200 genes (25-30%), and gaining less than 100 compared to the ancestor of all Bacilli*. Thus, the origin of Lactobacillales clearly involved major genome reduction. Table 2 lists some of the biologically interesting genes and genetic systems which apparently were lost or gained at the base of the Lactobacillales branch. 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 and one of the key enzymes of gluconeogenesis

(fructose-1,6-bisphosphatase) were lost; conversely, a variety of peptidases were acquired, apparently, via HGT. It is also likely that the Lactobacillales ancestor already was 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 non-orthologous gene displacements via HGT also were identified (Table 2), e.g., the DAHP synthase (the initial step of aromatic amino acid biosynthesis) typical of Gram-positive bacteria was displaced by the corresponding enzyme from Gram-negative bacteria. Furthermore, Lactobacilli apparently acquired via HGT, possibly, from an archaeal source, the complete melavonate pathway and lost the deoxyxylulose pathway of isoprenoid biosynthesis; the former pathway is common in archaea and rare in bacteria whereas the latter one is found in most bacteria. The acquisition of the mevalonate pathway at the base of the Lactobacilli tree, with a subsequent duplication of the mevalonate kinase gene, is supported by the organization of the genes for all enzymes of this pathway (two mevalonate kinases, mevalonate pyrophosphate decarboxylase, and isopentenyl-diphosphate delta-isomerase) in a single operon which is conserved in most lactobacillar genomes (see the “Genome Alignments” in the Supplemental Material).

Although genome reduction is the prevailing trend in the evolution of Lactobacilli, they also, like all other bacterial lineages {Jordan, 2001 #419}, have a significant number of expanded gene families which evolved either by lineage-specific gene duplication or by acquisition of paralogous genes via HGT {Koonin, 2001 #47}. A closer examination of these families is compatible with the notion that adaptation to growth in nutrient-rich environments had been the major driving force behind the fixation of duplications during the evolution of the Lactobacilli (suppl. “**Expansions”**). An interesting case of ancient gene acquisition is the second enolase which is characteristic of the Lactobacilli. All other bacteria have a single copy of this nearly ubiquitous glycolytic enzyme but Lhe lactobacilli have two. Phylogenetic tree analysis shows that one of these is the ancestral version in Gram-positive bacteria whereas the other one had been acquired by the ancestor of the Lactobacilli from a different bacterial lineage, most likely, actinobacteria. The evolution rate of both enolases seems to be increased in the Lactobacillar branches of the phylogenetic tree (Supplementary figure **4S “Enolase”**). The functional distinction between the two enolases could not be predicted from genome comparison; experimental identification of their functions should shed new light on the physiology of lactobacilli. Many other genes for proteins involved in sugar metabolism and transport were duplicated early in the evolution of the Lactobacilli, including PTS-systems, β-galactosidase, GpmB family sugar phosphatases, galactose mutarotase, L-lactate dehydrogenases of two distinct classes, and others. In addition to the apparent acquisition of new peptidases via HGT, we also observed several lineage-specific duplications of genes for these enzymes as well as those coding for amino acid transporters.

Several paralogous expansions include genes for proteins related to those involved in antibiotic resistance in other bacteria, such as β-lactamases and penicillin V acylase. However, all LAB species analyzed here have been shown to be sensitive to common antibiotics and, accordingly, "generally recognized as safe" (GRAS) for human consumption {Katla, 2001 #449; Teuber, 1999 #450}. Conceivably, the homologs of antibiotic resistance genes are involved in regular cell wall biosynthesis in the

Lactobacilli. 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 (PASTA) domains, are likely be important for regulation of cell wall biosynthesis {Yeats, 2002 #394}. Furthermore, Lactobacilli encode a paralog of class II lysyl-tRNA synthetase which is fused to a membrane-associated domain (COG2898) implicated in oxacillin-like antibiotic resistance in *Staphylococcus aureus* {Nishi, 2004 #420} and, probably, involved in cell wall biosynthesis.

Several Lactobacilli have multiple copies of competence-related genes which could enhance the potential of these bacteria for DNA uptake and HGT. Other expansions of paralogs awaiting experimental characterization include two RNA cytosine-C5-methylases (one of which is fused to another uncharacterized conserved domain, COG3270), three forms of ribosomal protein L33, which are predicted to be differentially regulated by zinc {Panina, 2003 #421}, two guanylate kinases, and several cation-transporting P-type ATPases.

The subsequent evolution of the Lactobacilli further enhanced the trends of genomic degradation and metabolic simplification. 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 biosynthesis of arginine and aromatic amino acids were lost independently in the *L. casei-L. delbruekii* group, in Lacbr, Pedpe and Oenoe; genes for serine and glycine biosynthesis have been lost in the common ancestor of Lactobacillales, and several fatty acid biosynthesis genes have been specifically lost in the Lacga and Lacjo branch (more examples are given **suppl. Table** "Biosynthesis"). Lineage-specific gene loss was extensive in the evolution of all lineages of Lactobacillales but several species were especially notable "losers". In particular, Strth not only lost numerous genes but also seems to have many fresh pseudogenes; similar findings have been recently reported for a different strain of the same species {Bolotin, 2004 #407}. Moreover, substantial gene loss (379 genes according to the present reconstruction) occurred also 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, MinD), in accord with the fact that this group includes a number of pathogenic bacteria {, 1995 #400}. Other lineages particularly prone to gene loss are Pedpe (489 genes lost), which has the smallest genome in the Lactobacillus group, and the Leume and Oenoe branch, with 380 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 *L. delbruekii* group (Lacbu, Lacga, Lacjo), leading to additional genome reduction in Lacga and Lacjo (**Fig. 3**).

Comparison between the number of genes lost or gained on a particular tree branch and the length of the corresponding branch reveals an intriguing pattern similar to that described previously by Snel and coworkers for proteobacteria {Snel, 2002 #118}. The number of gene losses (normalized by the size of the ancestral genome or not) strongly and significantly correlates with the branch length determined from sequence divergence (R=0.68, $p<5\text{x}10^{-4}$), 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 suggests evolution under purifying selection, occurring in small increments

along the evolutionary path; in contrast, gene gain appears to involve relatively large batches of genes acquired episodically, perhaps, under control of positive selection.

A lactic-acid-producing genome from the actinobacterial branch

In addition to the Lactobacillales genomes, we report the genome sequences of *B. longum* (Biflo) and *B. linens* (Breli) which belong to Actinobacteria, a high-GC lineage of Gram-positive bacteria that is only distantly related to Firmicutes {Ventura, 2004 #403; Rattray, 1999 #417}. On the basis of its phenotype, Biflo can be considered a lactic acid bacterium {, 1995 #400}. In contrast, Breli is not a lactic acid bacterium but is widely used in cheese-making for producing specific flavors {Rattray, 1999 #417}. Therefore it is interesting to compare these genomes to those of Lactobacilli, in an attempt to identify unique genomic features of lactic acid bacteria. To this end, we assigned all proteins of these two species to the LaCOGs for the purpose of comparison to Lactobacilli, and to the main set of COGs for the comparison to other fully sequenced actinobacteria (**Fig. 2**). As expected, the gene content of these bacteria is much more completely covered by the full COG set than by LaCOGs (76% vs 66% for Biflo and 77% vs 61% for Breli).

Biflo is the smallest actinobacterial genome sequenced to date, due to the substantial gene loss in this lineage. Among the 491 COGs that have been apparently lost by the Biflo lineage after the divergence from the common ancestor of actinobacteria, there are numerous genes that have been lost also by the common ancestor of Lactobacillales, including all subunits of Heme/copper-type cytochrome/quinol oxidase, Multisubunit Na+/H+ antiporter, Glycine cleavage system, heme biosynthesis, and others. Moreover, several gene gains in Biflo are also shared with the Lactobacillales, namely, phosphoketolase, the key enzyme for heterofermentation, numerous genes for sugar transport and catabolism, aminopeptidase C, and dipeptidase.

Among the 88 genes that are unique to the gastrointestinal symbionts Lacga and Lacjo, which are phenotypically similar to Biflo, only uncharacterized gene was found in Biflo (absent in other actinobacteria); in contrast, among the 12 genes specifically lost in this group, 8 were also missing in Biflo (6 of these genes are involved in fatty acid biosynthesis).

Since there are only 11 proteins from Biflo that were specifically shared with Lactobacilli to the exclusion of other bacteria (i.e., absent from COGs but included in LaCOGs), it appears that there is no unique gene set defining the lactic acid bacterial phenotype and most if not all metabolic capabilities of the Lactobacilli are encoded in genes shared by various bacterial lineages.

In a sharp contrast to Biflo, Breli almost fully retained all biosynthetic capabilities characteristic of free-living actinobacteria (suppl. Table **"Biosynthesis"**). In addition, Breli has 242 genes that probably have not been inherited from the actinobacterial ancestor but acquired later, including 17 genes shared with Biflo, Breli, and the Lactobacilli. These genes encode enzymes of energy, sugar, and amino acid metabolism (Na+/sugar permease, galactose mutarotase, dipeptidyl aminopeptidase) which seems relevant to the similar nutritional requirements of these bacteria. Thus, comparative-genomic analysis suggests considerable amount of HGT between phylogenetically distant bacteria which share habitats.

Phyletic patterns and central metabolism reconstruction

Given the specific phenotype of the Lactobacillales species, we were particularly interested in the sugar metabolism and energy conversion systems; we examined the evolution of these systems through phyletic patterns, i.e., patterns of presence-absence of genes in individual genomes {Koonin, 2003 #379}. The analysis of phyletic patterns for genes involved in these functions immediately shows that the most common pattern consists of genes represented in all species (**Fig. 4**). These include genes coding for the downstream part of glycolysis, from glyceralhehyde-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 PTS system. Clearly, these enzymes are insufficient to completely define the metabolism of any individual species; several reactions are specific to individual lineages (**Fig. 4)**. It is also notable that about half of the metabolic enzymes that are conserved in all Lactobacilli species are absent in Biflo, suggesting that this actinobacterium has very different pathways leading to the same end products. The presence/absence patterns of key enzymes of homo- and heterofermentation poorly correlate with the phenotypes of the Lactobacilli (suppl. Table "**Key enzymes**"). However, it has been shown that, under certain conditions, Lactobacilli switch from one type of fermentation to the other {Hemme, 2004 #399; Liu, 2003 #413}.

The metabolic potential of the Lactobacilli is complemented by the predicted transport capabilities. In particular, amino acid uptake systems dominate over sugar and peptide uptake systems, and among the detected sugar uptake systems, those specific for oligosaccharides and glycosides outnumber those for free sugars. In addition, Lactobacilli encode a variety of predicted drug, peptide, and macromolecular efflux pumps, some of which are likely to be involved in intercellular signaling. The transporter repertoire of the Lactobacilli is described in detail elsewhere {Lorca, 2005 #451}.

Other metabolic capabilities of Lactobacillales are listed in **suppl. Table "Biosynthesis".** Generally, Lacbr, Lacjo, Lacga, Lacde, and Pedpe have extremely narrow repertoires of biosynthetic pathways, whereas Lacla, Laccr, Breli, Lacpl, Leume retained the maximum number of biosynthetic reactions.

Specific features of individual lineages and species

Analysis of phyletic pattern frequencies gives a general picture of evolutionary trends characteristic of a selected set of organisms {Koonin, 2003 #379}. Among the most common non-trivial patterns, there are 44 LaCOGs (Lacca and Lacpl), 33 LaCOGs (Lacca, Lacpl and Lacbr) and 27 LaCOGs (Lacpl, Lacca, Pedpe, Lacbr) that reflect specific similarities in the gene content of Lacca and *Pediococcus* group (**Suppl. Table: "Frequent patterns"**). Interestingly, integrative approach to the phylogeny of LAB {Vandamme, 1996 #391} placed Lacca into the *Pediococcus* group and, specifically, into the *L. casei-Pediococcus* group. Among the less frequent but notable phyletic patterns are 18 LaCOGs shared by six species (Lacla, Laccr, Strth, Lacca, Lacpl, Leume), most of which reflect the ability of these organisms to synthesize tryptophan and histidine; 12 LaCOGs shared by all species except for Lacga and Lacjo, which is mostly due to the loss of fatty acid biosynthesis genes by the latter species; and 11 LaCOGs are present in all species other than Lacbr, Lacga and Lacjo which have lost the genes for purine biosynthesis enzymes (**Suppl. Table: "Frequent patterns"**).

Only a few Lactobacilli species have unique (i.e, missing in al other species analyzed here) biological systems (Suppl. Table: "**Unique features**"). Thus, Strth apparently has acquired a suit of genes for components of a predicted thermophile-specific repair system {Makarova, 2002 #96} and the 8 subunits of urease, an activity that has been experimentally characterized in this organism {Mora, 2004 #422}. Lacbr has a unique, complex system (18 genes, of which 17 is unique, comprising a single operon) for utilization of propanediol which has been experimentally studied mainly in *Salmonella enterica* {Bobik, 1999 #453} and seems to be linked to glycerol metabolism (a wine spoilage pathway) in *Lactobacillus collinoides* {Sauvageot, 2000 #454}. Pedpe encodes a Mn-dependent catalase, a key enzyme of peroxide degradation, which is not found in other lactic acid bacteria but has been biochemically characterized in Pediococcus {, 1995 #400}.

Lactobacillales are famous for producing specific antimicrobial peptides, the bacteriocins {Twomey, 2002 #405; Nes, 2004 #404}. Several proteins are responsible for the modification and export of bacteriocins, and regulation of bacteriocin biosynthesis and are often encoded in the same operon with the bacteriocin genes {Twomey, 2002 #405; Nes, 2004 #404}. Since bacteriocins are small proteins with highly diverged sequences, they are often hard to identify by amino acid conservation such that genome context analysis is required for a more complete characterization of the bacteriocin repertoire. Among the lactobacillar genomes analyzed here, 7 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 Pedpe, homologs of which are present also in Leume and Lacca (LaCOG01709). The second family consists of previously unnoticed putative bacteriocin precursors distantly related to Divercin V41 {Metivier, 1998 #423} and present in Pedpe and Lacjo (LaCOG03352). In addition, numerous small ORFs located in the immediate vicinity of the genes for bacteriocins and associated proteins might encode novel bacteriocins despite the lack of sequence similarity to known ones (Suppl. Figure **5S**). 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 (Suppl. Figure 5S).

Concluding remarks

To our knowledge, this work is the most extensive comparative analysis of a compact group of relatively closely related prokaryotic genomes which show a gradient of sequence conservation. Reduction of genome size and metabolic simplification are the central trends of LAB evolution. Major gene loss occurred already at the stage of a common ancestor of all Lactobacilli which indicates early adaptation to nutritionally rich environments. However, genome degradation appears to be an ongoing process: all species of Lactobacillales show loss of specific genes and many are enriched in pseudogenes. Beyond gene loss, lactobacilli have clear ancestral adaptations for nutritionally rich, microaerophilic environment which include duplication and acquisition via HGT of various enzymes of sugar and amino acid metabolism. Molecular systems responsible for the production of specific antimicrobials, such as the bacteriocins, are among other adaptations that become apparent through comparative-genomic analysis

and probably reflect the longterm existence of Lactobacilli in complex microbial communities. Comparison of the genomes of Lactobacilli with those of milk-digesting actinobacteria suggests that the LAB 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.

Comparative-genomic analysis described here also suggests a revision of the taxonomy of the Lactobacilli. Phylogenetic analysis of multiple protein sequences showed that the Streptococci-Lactococci branch is basal in the Lactobacillales tree, and the Pediococcus group is a sister to the Leuconostoc group, which supports the paraphyly of the Lactobacillus genus. Furthermore, *L. casei* is confidently placed at the base of the *L. delbruekii* group, which contradicts the earlier classification.

Acknowledgements:

This work was supported by Fidelity Systems, Inc., the US Department of Energy's Office of Science, Biological and Environmental Research Program and the by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36. Additional funding came from the American Vineyard Foundation (DAM), the CA Competitive Grants Program for Research in Viticulture and Enology (DAM), Western Dairy Center (BW), Utah Agricultural Experiment Station (BW), Dairy Management, Inc. (BW; TRK; DO), Danisco, Inc. (TRK), Southeast Dairy Foods Research Center (TRK), NC Dairy Foundation (TRK), NC Agricultural Research Service (TRK), NIH RO1 GM55434 (MS), the US Department of Agriculture, Agricultural Research Service, Food Science Research Unit, Raleigh, NC (FB), the Pickle Packers International, Inc. (FB), the Utah State Community-University Research Initiative (JB), the NE American Dairy Association (RH), the MN Agricultural Experiment Station project 18-062 (LM), the Minnesota South Dakota Dairy Foods Research Center (DO), the College of Agricultural and Life Science at the University of WI-Madison (JS).

Material and Methods:

Genomic DNA preparation.

Genomic DNA was prepared for each bacterium using standard protocols for LAB {Stahl, 1990 #426; Walker, 1994 #424}.

Shotgun sequencing.

Shotgun sequencing was carried out at the Joint Genome Institute (JGI). JGI randomly sheared the genomic DNA and blunt-end repaired the fragments using T4 polymerase and Klenow fragment. Fragments were size selected by agarose gel electrophoresis, a band cut out in the range of approximately 3Kb, purified from the gel and ligated into pUC18. Ligations were transformed into *E. coli* DH10B cells and colonies were picked into 384-well glycerol stocks. These subclone inserts were sequenced from both ends using universal primers. Each genome was sequenced to approximately 8X depth based on the estimated genome size and assembled using Jazz, the JGI assembler. Nearly 773 thousand lanes with an average pass rate (>50 Q20 bases per lane) of 90%. Produced over 400 Mb of data for the 11 genome projects.

Gap closure.

Gap closure was carried out at Fidelity Systems Inc.

Genome annotation and analysis

Open reading frames (ORFs) were identified using the GeneMarkS program {Besemer, 2001 #445}. Additionally, the remaining intergenic regions were compared to the non-redundant protein sequence database at the NCBI (http://www.ncbi.nlm.nih.gov) using the BLASTX program {Altschul, 1997 #43} and those ORFs that had detectable homologs (E-value <10-3, followed by manual validation of the alignments) were added to the list of predicted protein-coding genes. Gene functions were predicted by assigning predicted genes to the Clusters of Orthologous Groups of proteins (COGs; http://www.ncbi.nlm.nih.gov/COG) using the COGNITOR method {Tatusov, 2003 #428} and by additional database searches using the PSI-BLAST program {Altschul, 1997 #43}. Transfer RNAs were predicted using the tRNAscan-SE program {Lowe, 1997 #446}.

Construction of Lactobacillales Clusters of Orthologous Groups.

Lactobacillales Clusters of Orthologous Groups (LaCOGs) were constructed using the previously described procedure {Tatusov, 1997 #63; Tatusov, 2001 #116}, based on triangles of the best-hit relationships formed by proteins from different genomes. LaCOGs with only two organisms were constructed using reciprocal best-hit relationships. Lineage-specific expansions in Lactobacillales species were identified essentially as described in {Jordan, 2001 #419}. LaCOGs were linked to prokaryotic

Clusters of Orthologous Groups of proteins (COGs) using the COGNITOR approach {Tatusov, 2003 #428}.

Phylogenetic reconstructions

Multiple alignments of protein sequences were constructed using the MUSCLE program {Edgar, 2004 #430}; sites with more than 33% gap content were removed. Least squares trees were constructed using the PROTDIST and FITCH programs of the PHYLIP package {Felsenstein, 1996 #139} and Maximum Likelihood trees were constructed using the TREEPUZZLE program {Schmidt, 2002 #432}. The trees were examined for the compatibility with the molecular clock assumption using the LINTRE program {Takezaki, 1995 #433}. Alignments of the protein-coding nucleotide sequences were produced on the basis of the corresponding protein sequence alignments; synonymous and non-synonymous substitutions were analyzed using the CODEML program of the PAML package {Yang, 1997 #434}.

Whole genome similarity reconstructions

Inter-genome distance matrices were constructed using several previously described measures, namely, median distance between predicted orthologs (defined as reciprocal best hits, {Grishin, 2000 #437; Wolf, 2001 #368}), similarity of the gene content (number of shared LaCOGs, normalized by the size of the smaller genome, {Korbel, 2002 #436} and similarity of the gene order (fraction of genes covered by aligned gene chains). Alignments of gene order were constructed using the LamarckN program as described previously {Wolf, 2001 #429}. Similarity dendrograms were constructed using the neighbor-joining algorithm (the NEIGHBOR program of PHYLIP, {Felsenstein, 1996 #139}).

Consistency of local molecular clock

The consistency of local molecular clock in individual LaCOGs was tested using the previously described approach {Novichkov, 2004 #435}. Specifically, Maximum Likelihood distances between aligned LaCOG sequences were estimated using the CODEML program of PAML {Yang, 1997 #434}; if paralogs were present, the minimum distance between the paralogs from a pair of species was used to represent the inter-species distance. The matrix of inter-species distances was compared to the matrix of distances computed for concatenated alignments of ribosomal proteins. The residual variance after a straight-line zero-intercept approximation was compared to the complete variance of the distances; LaCOGs for which this residual variance was greater than the complete variance were considered to be the cases of severe deviation from the local molecular clock (as defined by the rate of evolution of the ribosomal proteins). For selected cases, the presence of HGT was verified by phylogenetic tree reconstruction.

Reconstruction of gene gains and losses

For the analysis of gene losses in the common ancestor of Lactobacillales, the complement of COGs present in at least one Lactobacillales species was compared to the COGs present in other Firmicutes. The COGs present in other Firmicutes but missing in Lactobacillales were considered lost by the common ancestor of Lactobacillales. The median number of genes per COG in Bacillales and Clostridiales was used as the estimate of the number of genes in the lost COGs. High and low estimates of the complement of COGs that were represented in the Lactobacillales and Bacilliales ancestor were obtained by requiring (for the low bound) or not requiring (for the upper bound) for a COG to be present in both Bacillales and non-Bacilli Firmicutes (Clostridiales or Mollicutes).

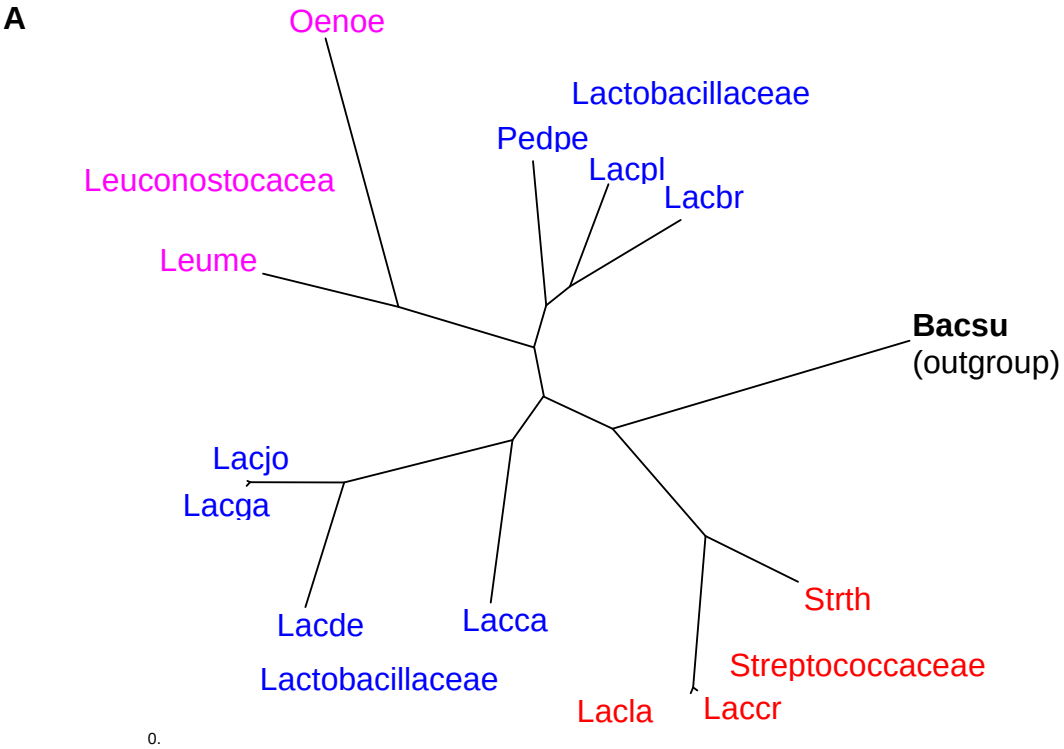
For the analysis of gene losses inside the Lactobacillales, the phyletic patterns of LaCOG were analyzed using a version of the weighted parsimony algorithm {Mirkin, 2003 #200} with the gain penalty of 2. A gene was assigned to the common ancestor of Lactobacillales and Bacillales if the members of a COG corresponding to a given LaCOG were present in at least one Bacillales species.

Figure legends

Figure 1. Phylogenetic trees of Lactobacilli constructed on the basis of concatenated alignments of ribosomal proteins (A) and RNA polymerase subunits(B). Abbreviations: Lacga - *Lactobacillus gasseri*, Lacbr - *Lactobacillus brevis*, Pedpe - *Pediococcus pentosaceus*, Laccr - *Lactococcus lactis cremoris*, Strth - *Streptococcus thermophilus*, Oenoe -*Oenococcus oeni*, Leume - *Leuconostoc mesenteroides*, Lacca - *Lactobacillus casei*, Lacde - *Lactobacillus delbrueckii*, Breli - *Brevibacterium linens*, Biflo - *Bifidobacterium longum*, Lacla - *Lactococcus lactis*, Lacpl - *Lactobacillus plantarum*, Lacjo - *Lactobacillus johnsonii*, Bacsu – *Bacillus subtilis*. All branches are supported at >75% bootstrap values. Species are colored according to the current taxonomy: Lactobacillaceae – blue; Leuconostocaceae – magenta; Streptococcaceae – red;

Figure 2. Conserved and unique genes in the genomes of Lactobacilli, *Bifidobacterium longum* and *Brevibacterium linens.*
"Homolog" indicate genes that have 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.
Abbreviations are as in **Fig.1. The** vertical axis shows the number of genes in a genome.

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



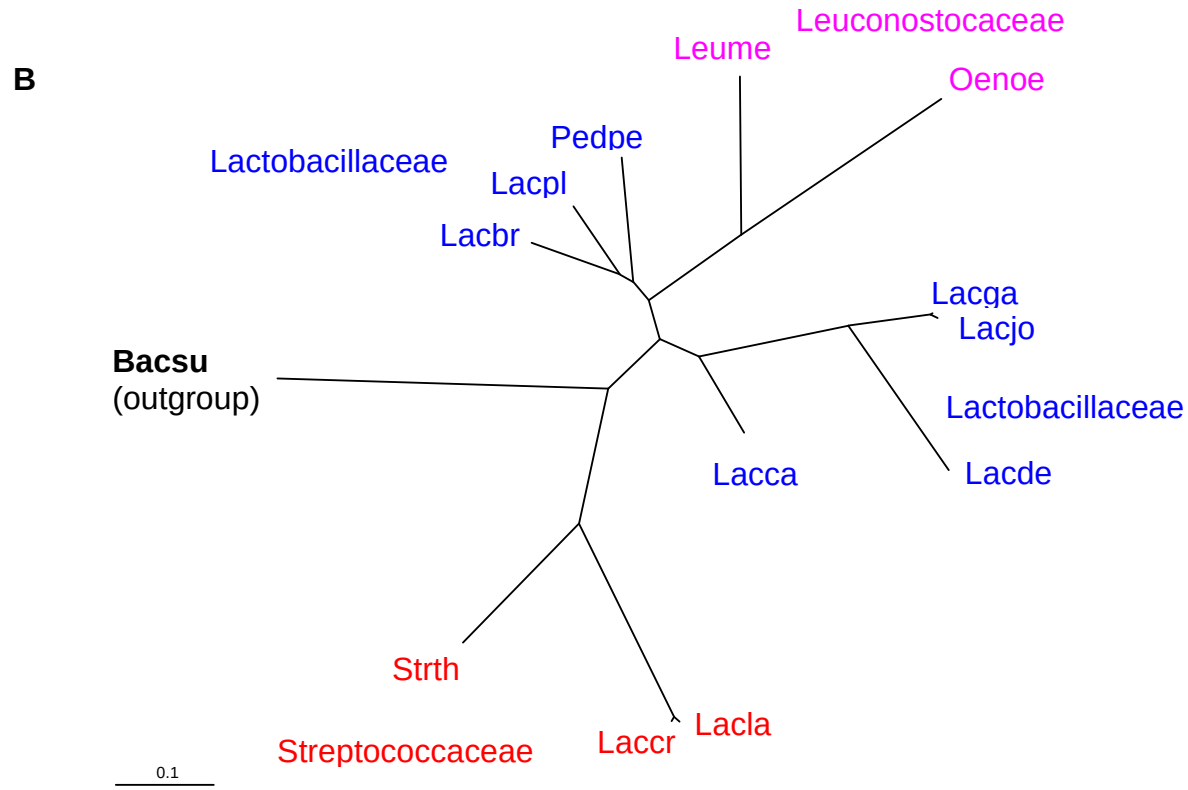


Figure 1A: Ribosomal proteins tree
Figure 1B: RNA polymerase subunits tree

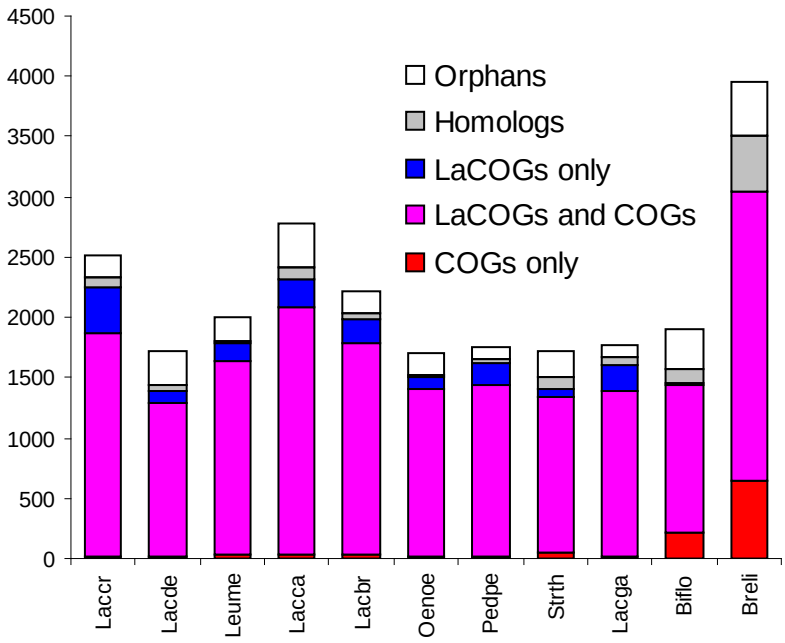


Figure 2. Conserved and unique genes.

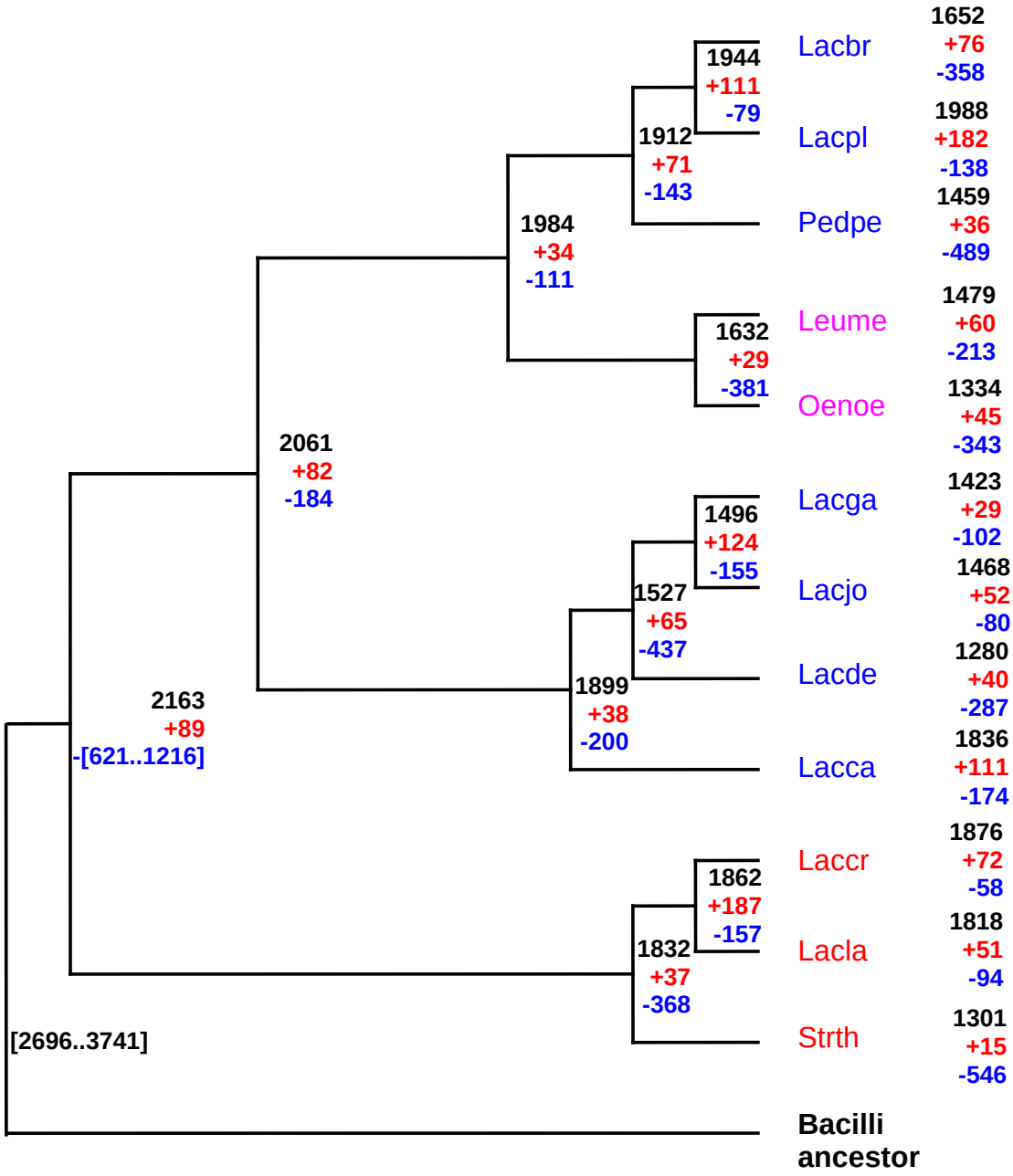


Figure 3. Gene loss and gain in evolution of lactobacilli

Fig. 4. 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: blue, *Lactobacillaceae*; magenta, *Leuconostocaceae*; red, *Streptococcaceae*.

Supporting Figure 5

Fig. 5. Proportion of genes (56%) that evolve similarly (same mode) with ribosomal proteins.

Supporting Figure 6

Fig. 6. Phylogenetic analysis of enolase. The maximum-likelihood unrooted tree was built with 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. Major branches of interest that were supported by bootstrap probability >70% are marked by black circles. The species analyzed in this work are shown in blue.

Supporting Figure 7
Fig. 7. A schematic representation of the reconstruction of key metabolic pathways associated with central carbon (carbohydrate) metabolism in lactic acid bacteria. Black arrows show reactions present in all species; pale blue arrows show reactions present in all species except one; and red arrows show reactions present 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 Table 7). In the phyletic patterns, "|" indicates the 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 gasseri*, *Lactobacillus delbrueckii*, and *Lactobacillus casei*. The systematic protein names (from *Escherichia coli*, *Bacillus 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 products of several reactions are dark green. An additional representation of the presence or absence of key genes associated with central carbon metabolism is listed in Table 7. +, Phyletic pattern for phosphoenolpyruvate carboxykinase, pckA, includes representatives of LaCOG2238 and a single protein from *Lb. casei*; *, acetoin reductase, ButA, homologs belong to the large family of short chain dehydrogenases from multiple LaCOGs. Many of these LaCOGs have unknown substrate specificity and might be involved in the same reaction as ButA; &, although there is no dedicated phosphotransferase system for lactose transport identified, it is possible that some phosphotransferase systems with wide substrate specificity can also transport lactose; ^, the presence of the system was essentially determined on the basis of the presence of this gene.

Supporting Figure 8

Fig. 8. Genome clusters encoding bacteriocins and genes for their export systems, including novel putative bacteriocins.

Table. 1

Species	Genome length, bp	Plasmids (no. of genes)	No. of proteins	No. of pseudogenes	No. of rRNA operons	No. of tRNAs	No. of prophages	Transposon-related ORFS, %
Lactobacillus gasseri	1,894,360	0	1,763	43	6	78	1	0.18
Lactobacillus brevis	2,340,228	pLVIS1 (12); pLVIS2 (25)	2,221	50	5	65	1	1.40
Pediococcus pentosaceus	1,832,387	0	1,757	19	5	55	2?	0.24
Lactococcus lactis ssp. cremoris	2,641,635	pLACR1 (11) pLACR2 (8) pLACR3 (64) pLACR4 (38) pLACR5 (8)	2,509	153	6	62	4?	4.82
Streptococcus thermophilus	1,864,178	pSTER1 (2) pSTER2 (4)	1,718	206	6	67	1?	3.72
Oenococcus oeni	1,780,517	0	1,701	120	2	43	0	0.54
Leuconostoc mesenteroides	2,075,763	pLEUM1 (34)	2,009	17	4	71	1	0.51
Lactobacillus casei	2,924,325	pLSEI1 (20)	2,776	82	5	59	2	3.25
Lactobacillus delbrueckiissp. bulgaricus	1,856,951	0	1,725	192	9	98	0	1.93

FIgure 4

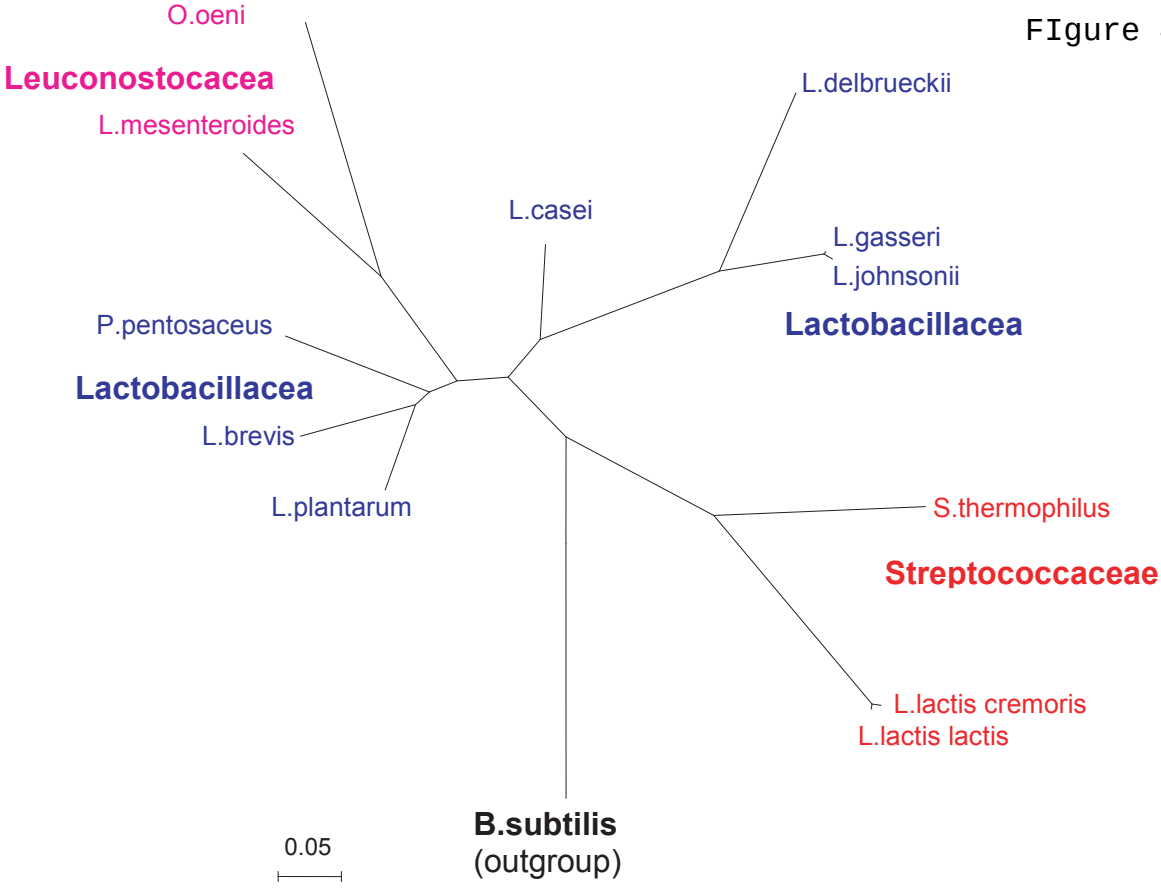

O.oeni
Leuconostocacea
L.mesenteroides
L.delbrueckii
L.casei
L.gasseri
L.johnsonii
Lactobacillacea
P.pentosaceus
Lactobacillacea
L.brevis
L.plantarum
S.thermophilus
Streptococcaceae
L.lactis cremoris
L.lactis lactis
B.subtilis
(outgroup)
0.05

Figure 5

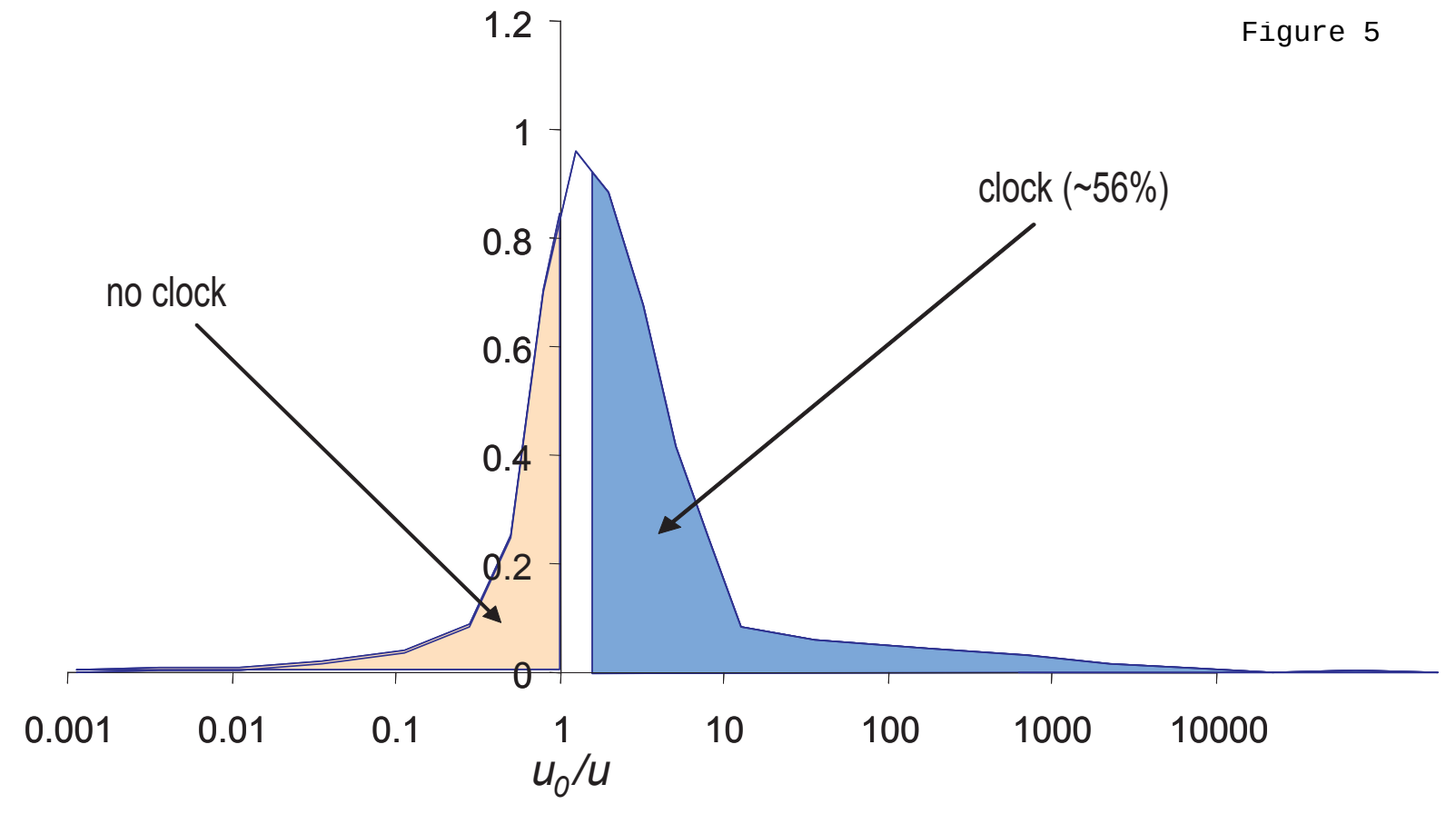

1.2
1
0.8
0.6
0.4
0.2
0
no clock
clock (~56%)
0.001
0.01
0.1
1
10
100
1000
10000
u_0/u

Table 3. Analysis of molecular clock violation in different functional groups of genes

Functional group code	Description of functional group	No. of COGs with u0/u*>1	No. of COGs with u0/u<=1	*P* (Chi^2)
X	Phages	14	33	0.0000
J	Translation, ribosomal structure and biogenesis	136	7	0.0000
G	Carbohydrate transport and metabolism	101	57	0.0002
L	Replication, recombination and repair	87	10	0.0020
V	Defense mechanisms	29	17	0.0336
Q	Secondary metabolites biosynthesis, transport and catabolism	7	6	0.0562
P	Inorganic ion transport and metabolism	51	25	0.0578
K	Transcription	93	40	0.0810
T	Signal transduction mechanisms	39	7	0.1785
F	Nucleotide transport and metabolism	66	14	0.1957
E	Amino acid transport and metabolism	132	32	0.2128
R	General function prediction only	138	34	0.2312
C	Energy production and conversion	42	18	0.2468
H	Coenzyme transport and metabolism	38	16	0.3008
U	Intracellular trafficking, secretion, and vesicular transport	12	2	0.4098
D	Cell cycle control, cell division, chromosome partitioning	16	3	0.4202
M	Cell wall/membrane/envelope biogenesis	78	20	0.4506
S	Function unknown	222	62	0.4715
I	Lipid transport and metabolism	37	13	0.6953
O	Posttranslational modification, protein turnover, chaperones	40	11	0.7269
N	Cell motility	4	1	0.8478
		u0/u>1, top 50%	**u0/u>1, bottom 50%**	
J	Translation, ribosomal structure and biogenesis	95	41	0.0000
G	Carbohydrate transport and metabolism	31	70	0.0001
X	Phages	0	14	0.0002
P	Inorganic ion transport and metabolism	13	38	0.0005
L	Replication, recombination and repair	57	30	0.0038
F	Nucleotide transport and metabolism	41	25	0.0489
I	Lipid transport and metabolism	24	13	0.0705
V	Defense mechanisms	10	19	0.0947
O	Posttranslational modification, protein turnover, chaperones	25	15	0.1138
K	Transcription	40	53	0.1776
C	Energy production and conversion	17	25	0.2170
U	Intracellular trafficking, secretion, and vesicular transport	8	4	0.2482
N	Cell motility	1	3	0.3173
D	Cell cycle control, cell division, chromosome partitioning	10	6	0.3173
H	Coenzyme transport and metabolism	16	22	0.3304
T	Signal transduction mechanisms	17	22	0.4233
M	Cell wall/membrane/envelope biogenesis	42	36	0.4969
E	Amino acid transport and metabolism	63	69	0.6015
S	Function unknown	108	114	0.6872
Q	Secondary metabolites biosynthesis, transport and catabolism	4	3	0.7055
R	General function prediction only	69	69	1.0000
	Total	691	691	

u, residual variance after local molecular clock approximation; u0, total variance of interspecies distances; u0/u<1, gross violation of local molecular clock.

Table 4. List of sugar metabolism genes that are prone to HGT

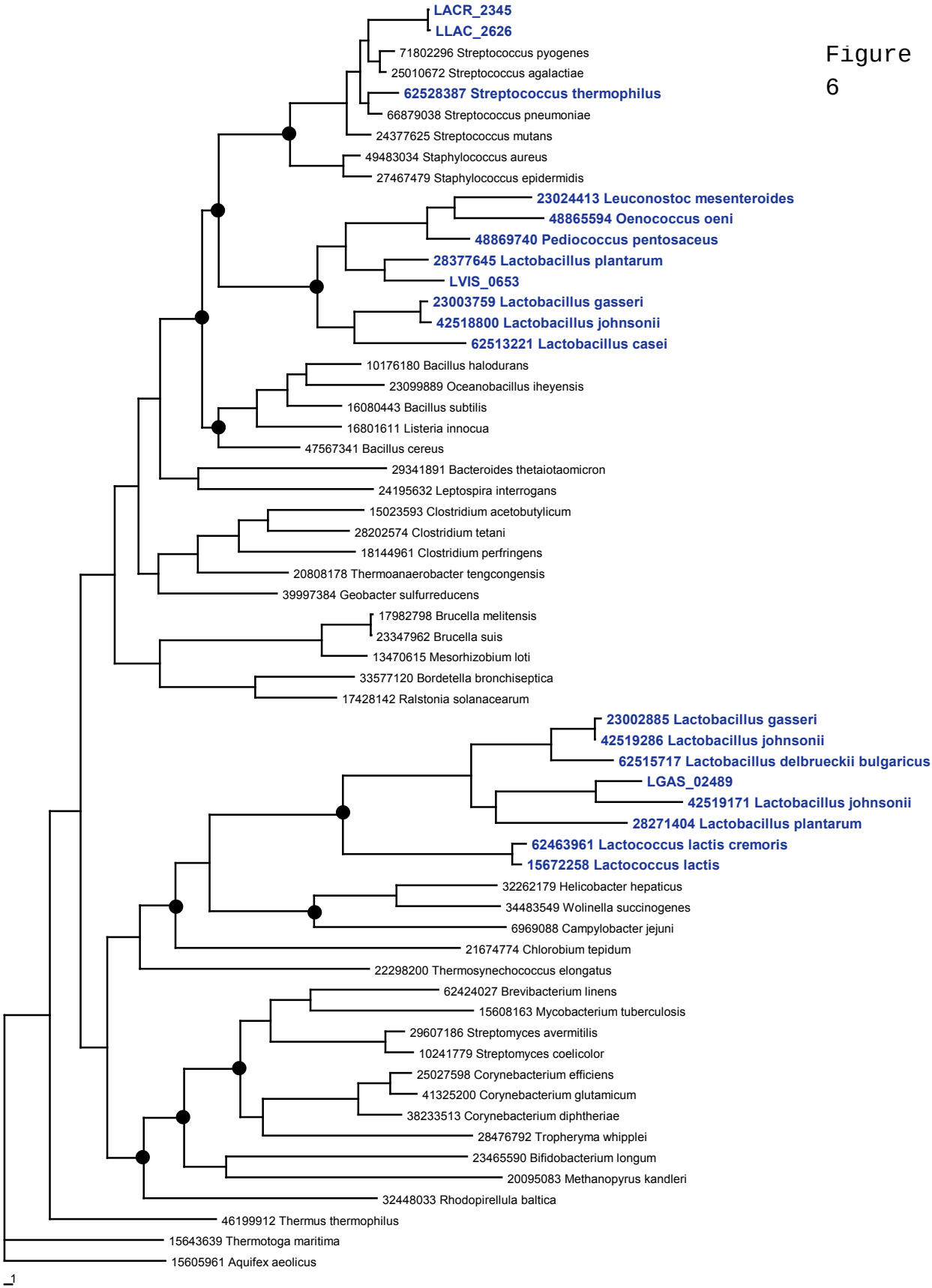
LaCOG number	LaCOG annotation
LaCOG00017	Mannitol/fructose-specific phosphotransferase system, IIA domain
LaCOG00047	Lactococcin A ABC transporter permease protein
LaCOG00075	Permeases of the major facilitator superfamily
LaCOG00178	Predicted sugar kinase
LaCOG00196	Predicted xylanase/chitin deacetylase
LaCOG00303	Cellobiose-specific PTS system IIB component
LaCOG00306	Beta-glucosidase/6-phospho-beta-glucosidase/beta-galactosidase
LaCOG00309	ABC-type sugar transport systems, ATPase components
LaCOG00313	Phosphotransferase system IIA components
LaCOG00343	Beta-glucosidase-related glycosidases
LaCOG00368	Glycerol-3-phosphatase transporter
LaCOG00469	Glucan phosphorylase
LaCOG00526	Phosphomannose isomerase
LaCOG00594	Glycerate kinase
LaCOG00634	Permeases of the major facilitator superfamily
LaCOG00792	Permeases of the major facilitator superfamily
LaCOG00804	Citrate lyase synthetase
LaCOG00805	Citrate lyase, gamma subunit
LaCOG00806	Citrate lyase beta subunit
LaCOG00807	Citrate lyase, alpha subunit
LaCOG00842	Glycerol uptake facilitator and related permeases (Major Intrinsic Protein Family)
LaCOG00913	Permeases of the major facilitator superfamily
LaCOG00948	Beta-glucoside-specific PTS system IIABC component
LaCOG00976	Phosphoketolase
LaCOG00979	Beta-xylosidase
LaCOG00980	Na+/xyloside symporter and related transporters
LaCOG01030	Glucosamine-6-phosphate isomerase
LaCOG01064	Transketolase
LaCOG01066	2-dehydro-3-deoxygluconokinase
LaCOG01073	Ribose/xylose/arabinose/galactoside ABC-type transport systems, permease components
LaCOG01105	Neopullulanase
LaCOG01108	ABC-type maltose transport systems, permease component
LaCOG01118	Mannose-specific PTS system component IIC
LaCOG01119	Mannose-specific PTS system component IID
LaCOG01123	Fructose-2,6-bisphosphatase
LaCOG01252	Permeases of the major facilitator superfamily
LaCOG01321	Beta-galactosidase
LaCOG01473	Gluconate kinase
LaCOG01474	6-phosphogluconate dehydrogenase
LaCOG01569	Fructose-2,6-bisphosphatase
LaCOG01608	Permeases of the major facilitator superfamily
LaCOG01697	Beta-glucosides PTS, EIIBCA
LaCOG01733	Uncharacterized conserved protein
LaCOG01797	Permeases of the major facilitator superfamily
LaCOG01813	Permeases of the major facilitator superfamily
LaCOG01909	Permeases of the major facilitator superfamily
LaCOG02055	Alpha-galactosidase
LaCOG02062	Phosphotransferase system mannitol/fructose-specific IIA domain (Ntr-type)
LaCOG02063	Galacitol PTS, EIIC
LaCOG02064	Phosphotransferase system, galactitol-specific IIB component
LaCOG02069	Sugar phosphate isomerases/epimerases
LaCOG02237	Glycosidases
LaCOG02279	Phosphotransferase system, mannose/fructose-specific component IIA
LaCOG02290	3-hexulose-6-phosphate synthase and related proteins
LaCOG02306	Phosphotransferase system, mannose/fructose/N-acetylgalactosamine-specific component IIB
LaCOG02307	Phosphotransferase system, mannose/fructose/N-acetylgalactosamine-specific component IIC
LaCOG02308	Phosphotransferase system, mannose/fructose/N-acetylgalactosamine-specific component IID

Table 5. Examples of gene loss and gain associated with the last common ancestor of *Lactobacillales*

COG functional group	Enzyme/system name	LaCOGs	Corresponding COG
Examples of genes/systems gained			
Sugar and energy metabolism	Phosphoenolpyruvate carboxylase, ppc	LaCOG02200	COG1892
	Citrate lyase, citCDFE	LaCOG00804	COG3053
		LaCOG00805	COG3052
		LaCOG00806	COG2301
		LaCOG00807	COG3051
Amino acid metabolism	Aminopeptidase N, pepN	LaCOG00215	COG0308
	Dipeptidase, pepD	LaCOG00171	COG4690
	Neutral endopeptidase, pepO	LaCOG01198	COG3590
	Pyrrolidone-carboxylate peptidase	LaCOG02151	COG2039
Cofactor biosynthesis	Nicotinamide mononucleotide transporter, pnuC	LaCOG00611	COG3201
	Bifunctional Nicotinamide ribose kinase/Nicotinamide mononucleotide adenylyltransferase, nadR	LaCOG01327	COG1056 COG3172
Transporters	K+ transporter, kup	LaCOG00418	COG3158
Lipid biosynthesis	3-hydroxy-3-methylglutaryl CoA synthase	LaCOG01029	COG3425
	Hydroxymethylglutaryl-CoA reductase	LaCOG01027	COG1257
	Mevalonate kinase	LaCOG00296	COG1577
	Phosphomevalonate kinase	LaCOG00298	COG1577
	Mevalonate pyrophosphate decarboxylase	LaCOG00297	COG3407
	Isopentenyl diphosphate isomerase	LaCOG00299	COG1304
Examples of Genes/Systems Lost			
Sugar and energy metabolism	Heme/copper-type cytochrome/quinol oxidase, CyoABCD		COG3125 COG1845 COG0843 COG1622
	Fructose-1,6-bisphosphatase;		COG0158
	Phosphoglyceromutase		COG0696
Amino acid metabolism	Glycine cleavage system, gcvTRP		COG0404 COG0403 COG1003
	Methionine synthase I (cobalamin-dependent), metH		COG0646
	Na+/alanine symporter, alsT		COG1115
	Na+/proline symporter, putP		COG0591

Cofactor biosynthesis	Most of heme biosynthesis, hemABCDEFLY		COG1232
			COG0635
			COG0001
			COG0407
			COG1587
			COG0181
			COG0113
			COG0373
	Molybdenum cofactor biosynthesis, moaABCDE, mobAB, moeA		COG0303
			COG1763
			COG0746
			COG0314
			COG1977
			COG0315
			COG0521
			COG2896
	Panthothenate biosynthesis, panBCD		
			COG0853
			COG0414
			COG0413

Figure 6


LACR_2345
LLAC_2626
71802296 Streptococcus pyogenes
25010672 Streptococcus agalactiae
62528387 Streptococcus thermophilus
66879038 Streptococcus pneumoniae
24377625 Streptococcus mutans
49483034 Staphylococcus aureus
27467479 Staphylococcus epidermidis
23024413 Leuconostoc mesenteroides
48865594 Oenococcus oeni
48869740 Pediococcus pentosaceus
28377645 Lactobacillus plantarum
LVIS_0653
23003759 Lactobacillus gasseri
42518800 Lactobacillus johnsonii
62513221 Lactobacillus casei
10176180 Bacillus halodurans
23099889 Oceanobacillus iheyensis
16080443 Bacillus subtilis
16801611 Listeria innocua
47567341 Bacillus cereus
29341891 Bacteroides thetaiotaomicron
24195632 Leptospira interrogans
15023593 Clostridium acetobutylicum
28202574 Clostridium tetani
18144961 Clostridium perfringens
20808178 Thermoanaerobacter tengcongensis
39997384 Geobacter sulfurreducens
17982798 Brucella melitensis
23347962 Brucella suis
13470615 Mesorhizobium loti
33577120 Bordetella bronchiseptica
17428142 Ralstonia solanacearum
23002885 Lactobacillus gasseri
42519286 Lactobacillus johnsonii
62515717 Lactobacillus delbrueckii bulgaricus
LGAS_02489
42519171 Lactobacillus johnsonii
28271404 Lactobacillus plantarum
62463961 Lactococcus lactis cremoris
15672258 Lactococcus lactis
32262179 Helicobacter hepaticus
34483549 Wolinella succinogenes
6969088 Campylobacter jejuni
21674774 Chlorobium tepidum
22298200 Thermosynechococcus elongatus
62424027 Brevibacterium linens
15608163 Mycobacterium tuberculosis
29607186 Streptomyces avermitilis
10241779 Streptomyces coelicolor
25027598 Corynebacterium efficiens
41325200 Corynebacterium glutamicum
38233513 Corynebacterium diphtheriae
28476792 Tropheryma whipplei
23465590 Bifidobacterium longum
20095083 Methanopyrus kandleri
32448033 Rhodopirellula baltica
46199912 Thermus thermophilus
15643639 Thermotoga maritima
15605961 Aquifex aeolicus
1

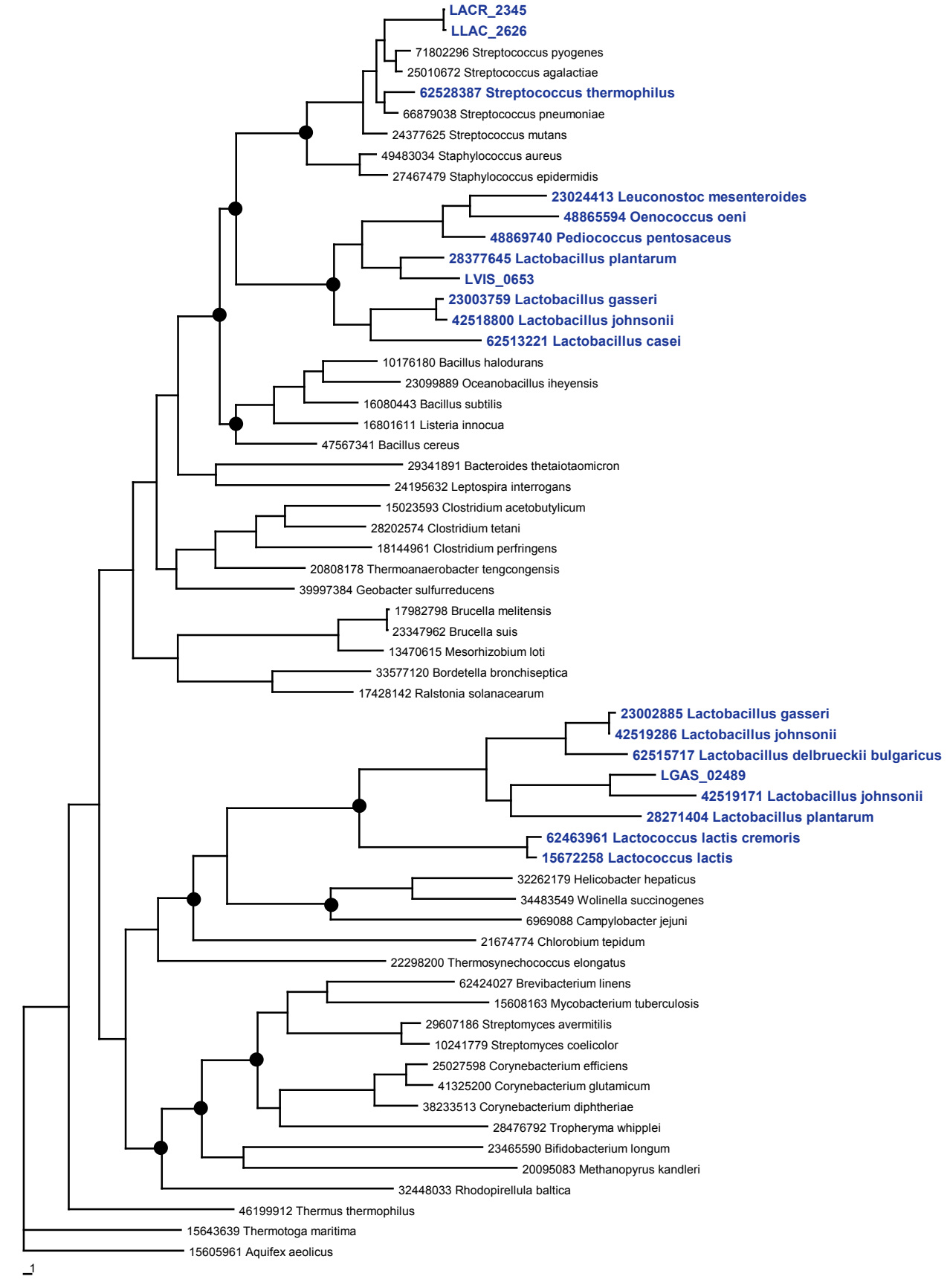
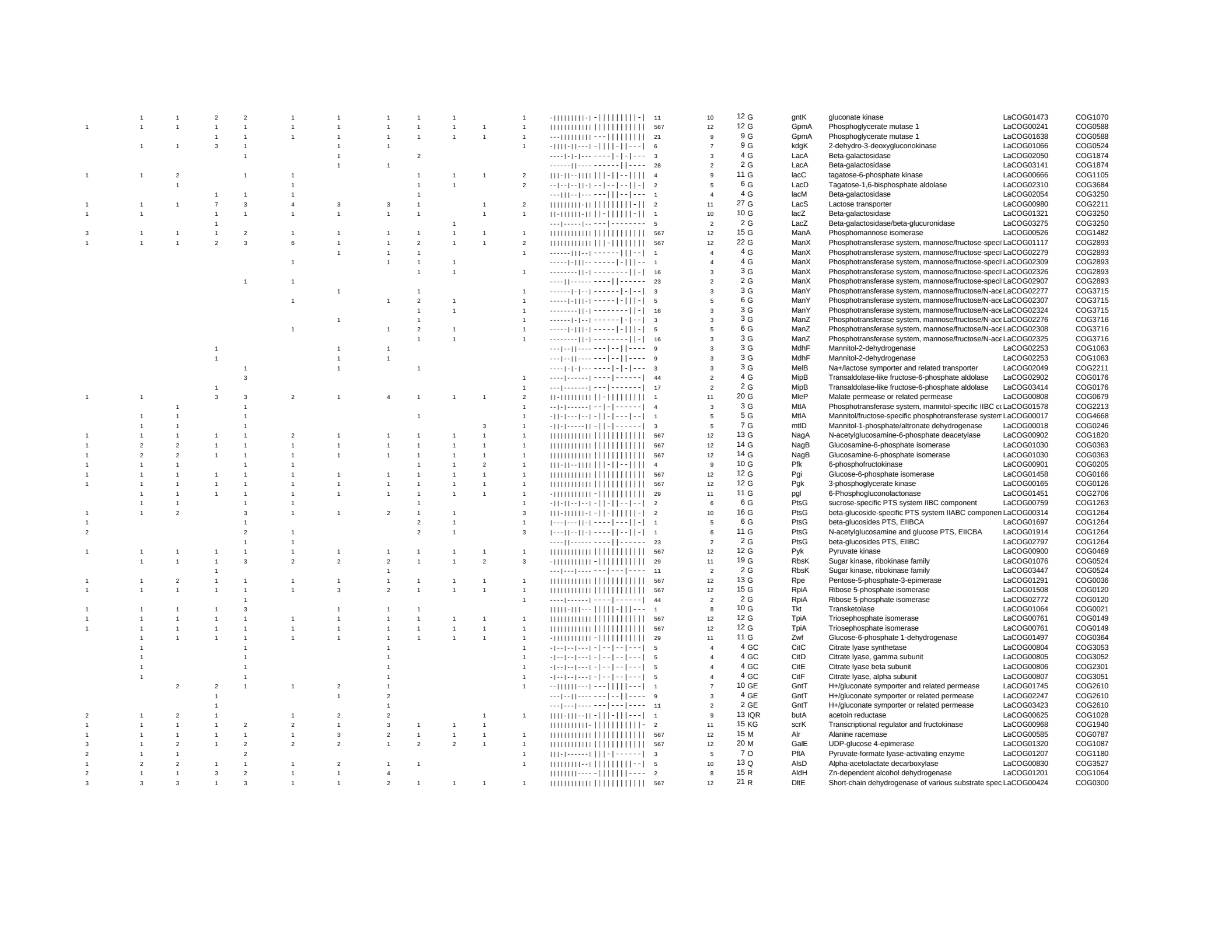
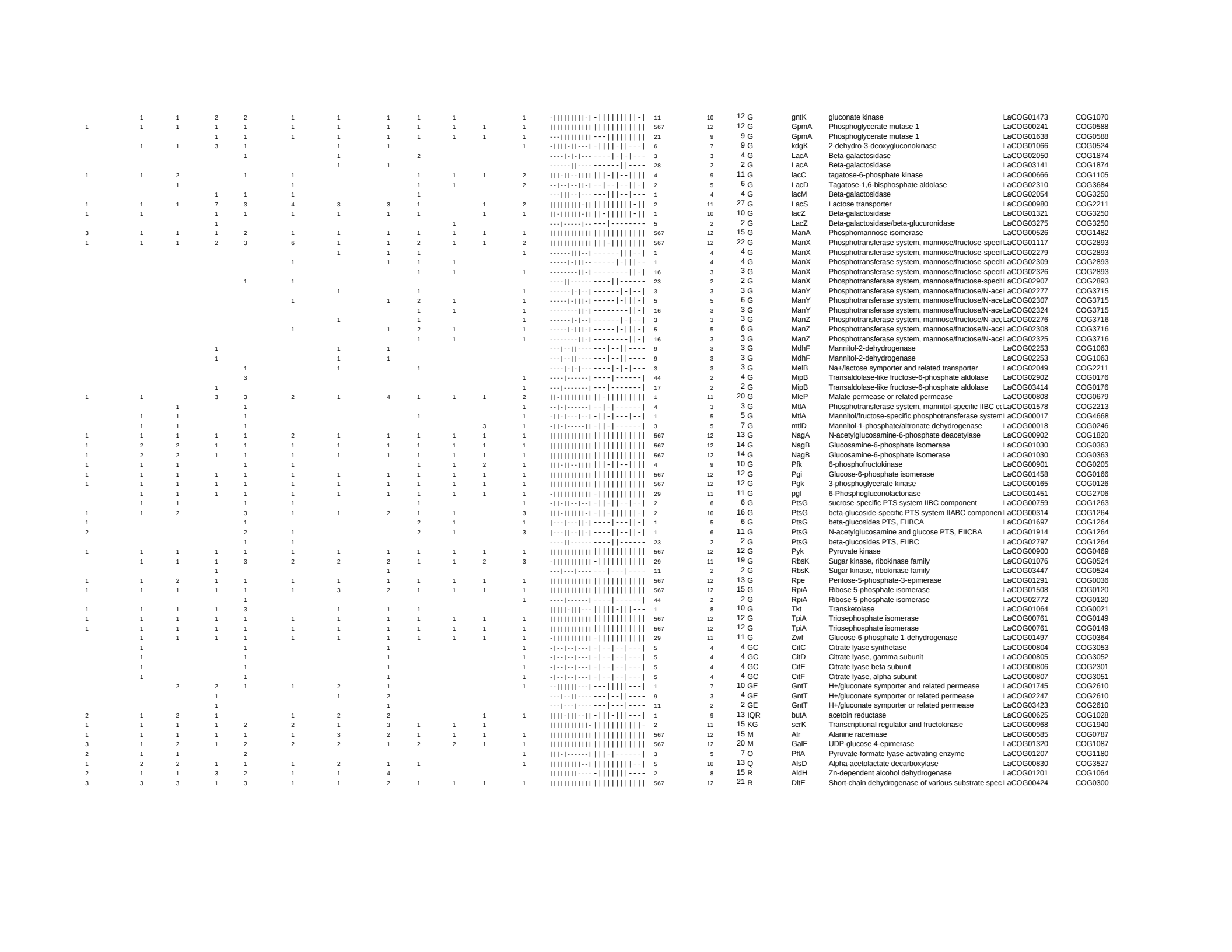

LACR_2345
LLAC_2626
71802296 Streptococcus pyogenes
25010672 Streptococcus agalactiae
62528387 Streptococcus thermophilus
66879038 Streptococcus pneumoniae
24377625 Streptococcus mutans
49483034 Staphylococcus aureus
27467479 Staphylococcus epidermidis
23024413 Leuconostoc mesenteroides
48865594 Oenococcus oeni
48869740 Pediococcus pentosaceus
28377645 Lactobacillus plantarum
LVIS_0653
23003759 Lactobacillus gasseri
42518800 Lactobacillus johnsonii
62513221 Lactobacillus casei
10176180 Bacillus halodurans
23099889 Oceanobacillus iheyensis
16080443 Bacillus subtilis
16801611 Listeria innocua
47567341 Bacillus cereus
29341891 Bacteroides thetaiotaomicron
24195632 Leptospira interrogans
15023593 Clostridium acetobutylicum
28202574 Clostridium tetani
18144961 Clostridium perfringens
20808178 Thermoanaerobacter tengcongensis
39997384 Geobacter sulfurreducens
17982798 Brucella melitensis
23347962 Brucella suis
13470615 Mesorhizobium loti
33577120 Bordetella bronchiseptica
17428142 Ralstonia solanacearum
23002885 Lactobacillus gasseri
42519286 Lactobacillus johnsonii
62515717 Lactobacillus delbrueckii bulgaricus
LGAS_02489
42519171 Lactobacillus johnsonii
28271404 Lactobacillus plantarum
62463961 Lactococcus lactis cremoris
15672258 Lactococcus lactis
32262179 Helicobacter hepaticus
34483549 Wolinella succinogenes
6969088 Campylobacter jejuni
21674774 Chlorobium tepidum
22298200 Thermosynechococcus elongatus
62424027 Brevibacterium linens
15608163 Mycobacterium tuberculosis
29607186 Streptomyces avermitilis
10241779 Streptomyces coelicolor
25027598 Corynebacterium efficiens
41325200 Corynebacterium glutamicum
38233513 Corynebacterium diphtheriae
28476792 Tropheryma whipplei
23465590 Bifidobacterium longum
20095083 Methanopyrus kandleri
32448033 Rhodopirellula baltica
46199912 Thermus thermophilus
15643639 Thermotoga maritima
15605961 Aquifex aeolicus
1

Table 7. Complete list of LaCOGs and their annotation related to the reconstruction of central carbohydrate metabolism (support for Table 8)

Note: The no. of protein-coding genes (including pseudogenes) for each genome is indicated for each LaCOG in the first 12 columns. In phyletic pattern schemes, "|" means presence and "-" means absence of a gene (genes) from the corresponding LaCOG. Functional group letters are described at www.ncbi.nlm.nih.gov/COG/old/palox.cgi?fun=all.

S. thermophilus	*Lc. lactis* ssp. *lactis*	*Lc. lactis* ssp. *cemoris*	*Lb. brevis*	*Lb. plantarum*	*P. pentosaceus*	*Le. mesenteroides*	*O. oeni*	*Lb. johnsonii*	*Lb. gasseri*	*Lb. delbrueckii* ssp. *bulgaricus*	*Lb. casei*	Phyletic pattern (presence/absence)	Phyletic pattern (pseudogenes are not included)	Pattern type count	No. of organisms	No. of proteins	COG functional group	COG gene	LaCOG annotation	LaCOG no.	Corresponding COG
1	2	2	3	3	2	2	1	2	2	1	3	|||||||||||	|||||||||||	567	12	24	C	ackA	Acetate kinase	LaCOG01348	COG0282
5	1	1	3	2		1	1	1	1	4	2	||||-||||||	-||||-||||||	15	11	22	C	adh	alcohol-acetaldehyde dehydrogenase	LaCOG01452	COG1454
1										1		|---------|-	----------	4	2	2	C	adh	alcohol-acetaldehyde dehydrogenase	LaCOG02084	COG1454
1	1	1	1	1	1	1	1	1	1	1	1	|||||||||||	|||||||||||	567	12	12	C	AtpA	F0F1-type ATP synthase, alpha subunit	LaCOG01170	COG0056
1	1	1	1	1	1	1	1	1	1	1	1	|||||||||||	|||||||||||	567	12	12	C	AtpB	F0F1-type ATP synthase, subunit a	LaCOG01173	COG0356
1	1	1	1	1	1	1	1	1	1	1	1	|||||||||||	|||||||||||	567	12	12	C	AtpC	F0F1-type ATP synthase, epsilon subunit	LaCOG01167	COG0355
1	1	1	1	1	1	1	1	1	1	1	1	|||||||||||	|||||||||||	567	12	12	C	AtpD	F0F1-type ATP synthase, beta subunit	LaCOG01168	COG0055
1	1	1	1	1	1	1	1	1	1	1	1	|||||||||||	|||||||||||	567	12	12	C	AtpE	F0F1-type ATP synthase, subunit c/Archaeal/vacuola	LaCOG01174	COG0636
1	1	1	1	1	1	1	1	1	1	1	1	|||||||||||	|||||||||||	567	12	12	C	AtpF	F0F1-type ATP synthase, subunit b	LaCOG01172	COG0711
1	1	1	1	1	1	1	1	1	1	1	1	|||||||||||	|||||||||||	567	12	12	C	AtpG	F0F1-type ATP synthase, gamma subunit	LaCOG01169	COG0224
1	1	1	1	1	1	1	1	1	1	1	1	|||||||||||	|||||||||||	567	12	12	C	AtpH	F0F1-type ATP synthase, delta subunit	LaCOG01171	COG0712
			1	1	1	1	1	1	1	1	1	---|||||||||	---|||||||||	21	9	9	C	FumC	Fumarase	LaCOG01721	COG0114
	1	1			1				1	1		-||-|---||--||-|---||--	-||-|---||--||-|---||--	2	5	5	C	GalT	Galactose-1-phosphate uridylyltransferase	LaCOG00577	COG1085
1	1	2										|||---------	||----------	69	3	4	C	GltA	Citrate synthase	LaCOG00448	COG0372
1	1	1	1	1	1			1	1	1	2	||||||--||||	||||||--||||	17	10	11	C	ldh	L-lactate dehydrogenase	LaCOG00899	COG0039
	1	1	2	1	1			1	1	1	3	-|||||--||||	-|---||--||--|-|	4	9	12	C	LldD	L-lactate dehydrogenase (FMN-dependent) and relate	LaCOG00845	COG1304
	1			1			1				1	-|--|--|---|	-|---|--|--|---|	5	4	4	C	mae	Malate oxidoreductase	LaCOG00802	COG0281
				1				1	1			----|---||--|----	----|---||--||--	12	3	3	C	Mdh	Malate/lactate dehydrogenase	LaCOG01720	COG0039
	1	1										-||---------	-||---------	246	2	2	C	Mdh	Malate/lactate dehydrogenase	LaCOG02549	COG0039
			1								1	---|-------|	---|-------|	17	2	2	C	Mdh	Malate/lactate dehydrogenase	LaCOG03372	COG0039
1	1	2		3	1	1	3	1	1	1	1	|||-||||||||	|||-||||||||	1	11	16	C	mdh/Idh	Enzyme with possible activities of L-2-hydroxyisocapr	LaCOG00274	COG0039
	1	1				1					1	-||---|----|	-||---|----|	8	4	4	C	mleP	Na+/malate symporter	LaCOG00628	COG3493
	1	3	1	1	1	1	1				1	-|||||||---|	-|-|||||---|	8	8	10	C	mleS	Malolactic enzyme	LaCOG00627	COG0281
				1				1	1	2	1	----|---||||	----|---||-||-|	7	5	6	C	PckA	Phosphoenolpyruvate carboxykinase (ATP)	LaCOG02038	COG1866
1	1	1	1	1	1	1	1				2	|||||||---|	|||||||---|	21	9	10	C	Pdh/AceF	acetoin/pyruvate dehydrogenase complex, E2 compo	LaCOG00031	COG0508
1	1	1	1	1	1	1	1				2	|||||||---|	|||||||---|	21	9	10	C	Pdh/AcoA	acetoin dehydrogenase complex, E1 component, alpl	LaCOG00033	COG1071
2	3	4	1	1	2	1	1	1	1	1	2	|||||||||||	|||||||||||	567	12	20	C	Pdh/Lpd	acetoin/pyruvate dehydrogenase complex, E3 compo	LaCOG00030	COG1249
1	1	1		2							1	|||-|------|	|||-|------|	3	5	6	C	PflD	Pyruvate-formate lyase	LaCOG00443	COG1882
1	1	1		2							1	|||-|------|	|||-|------|	3	5	6	C	PflD	Pyruvate-formate lyase	LaCOG00443	COG1882
						1	1					------||----	------||----	28	2	2	C	Ppc	Phosphoenolpyruvate carboxylase (archaeal type)	LaCOG01680	COG1892
1								2	1	1		|-------|||-	|-------|||-	2	4	5	C	Ppc	Phosphoenolpyruvate carboxylase	LaCOG02200	COG2352
1	1	1	1	1	1	1	1	1	1	1	2	|||||||||||	|||||||||||	567	12	13	C	Pta	Phosphotransacetylase	LaCOG01092	COG0280
	1	1		1	1						2	-||-||-----|	-||-||-----|	10	5	6	C	Pyc	Pyruvate carboxylase	LaCOG00447	COG1038
	1	1		2	1	1		1	1	6	1	-||-|||-||||	-||-||-||||	1	9	15	C	SdhA	Fumarate reductase, flavoprotein subunit	LaCOG00754	COG1053
				3					2			----|----|-- ----|----|--		2	2	5	C	SdhA	Succinate dehydrogenase/fumarate reductase, flavop	LaCOG02805	COG1053
			2	2	1	4	4	1	1	6	2	|||||||||||	---|||||||--	567	9	23	CHE	Ldh	Lactate dehydrogenase or related 2-hydroxyacid deh	LaCOG00403	COG1052
2	1		1		1						1	||-|-|-----|	-|-|-|-----|	1	4	5	E	Ald	Alanine dehydrogenase	LaCOG00837	COG0686
1	1	1	1	1	1	1	2	1	1		1	|||||||||-|	|||||||||-|	21	11	12	EH	IlvB	Acetolactate synthase	LaCOG00801	COG0028
1	1	1				1						|||---|-----	|||---|-----	13	4	4	EH	IlvB	Acetolactate synthase large subunit	LaCOG00826	COG0028
							1				1	-------|---|	-------|---|	20	2	2	EH	IlvB	Thiamine pyrophosphate-requiring enzymes [acetolac	LaCOG03416	COG0028
	2	2	1	4	2	1	1	1	1	5	1	-||||||||||	-|-||||||||	29	11	21	EH	Pox	Pyruvate oxidase or other thiamine pyrophosphate-re	LaCOG01367	COG0028
				1							1	----|------|	----|------|	44	2	2	EH	Pox	Pyruvate oxidase	LaCOG02904	COG0028
2	1	2	1	1	1	1	1	1	1	1	2	|||||||||||	||-|||||||	567	12	15	G	AraH/RbcC	Ribose/xylose/arabinose/galactoside ABC-type transp	LaCOG01073	COG1172
	1	1							1		1	-||-----|--|	-||-----|--|	2	4	4	G	BglB	Beta-glucosidase/6-phospho-beta-glucosidase/beta-g	LaCOG00114	COG2723
5	4	7	1	9	4	3	4	3	6	3	3	|||||||||||	-||||||||||	567	12	52	G	BglB	Beta-glucosidase/6-phospho-beta-glucosidase/beta-g	LaCOG00306	COG2723
	1	1		1	1		1	1	1			-||-||-|||--	-||-||-|||--	1	7	7	G	BglB	Beta-glucosidase/6-phospho-beta-glucosidase/beta-g	LaCOG00568	COG2723
	1	1	1								2	-|||-------|	-|||-------|	5	4	5	G	Eda	2-keto-3-deoxy-6-phosphogluconate aldolase	LaCOG01065	COG0800
	1	1		1				2				-||-|---|||-	-||-|---|||-	3	6	9	G	Eno	Enolase	LaCOG00189	COG0148
1	2	1	1	1	1	1	1	1			1	|||||||||-|	|||||||||-|	21	11	12	G	Eno	Enolase	LaCOG00434	COG0148
1	1	1		1	1		2	1	1	1	3	|||-||-||||	|||-||-||||	2	10	13	G	Fba	Fructose/tagatose bisphosphate aldolase	LaCOG01278	COG0191
	1		1	2	1	2	2	1	1	1	1	-|-||||||||	-|-||||||||	3	10	13	G	Fpk/Xpk	Phosphoketolase	LaCOG00976	COG3957
					1			1	1		1	-----|--||-|	-----|--||-|	3	4	4	G	FruK	Fructose-1-phosphate kinase or related fructose-6-ph	LaCOG02311	COG1105
								1	1			--------||--	--------||--	88	2	2	G	FruK	Fructose-1-phosphate kinase or related fructose-6-ph	LaCOG03212	COG1105
1	1	1	1	1	2	3	1	1	1		1	|||||||||-|	|||||||||-|	21	11	14	G	GalK	Galactokinase	LaCOG01324	COG0153
1	2	2	3	3	3	2	2	1	1		2	|||||||||-|	|||||||||-|	21	11	22	G	GalM	aldose 1-epimerase	LaCOG00981	COG2017
	1	1	1		1			1	1		1	-|||-|--||-|||-|--||-|	-|||-|--||-|||-|--||-|	1	7	7	G	GalM	Galactose mutarotase and related enzymes	LaCOG01068	COG2017
	1	2	1	1	1	1	1	1	1	1	1	-||||||||||	-||||||||||	29	11	12	G	GalM	Galactose mutarotase or related enzyme	LaCOG01288	COG2017
		2							1	2	1	--|------|||	--|------|||	1	4	6	G	GalpI	Galactose 6-phosphate isomerase	LaCOG02106	COG0698
		1							1		1	--|------|-|	--|------|-|	1	3	3	G	GalpI	Galactose 6-phosphate isomerase	LaCOG03238	COG0698
1	1	1	1	1	1	2	1	1	1		1	|||||||||-|	|||||||||-|	21	11	12	G	GalT	Galactose-1-phosphate uridyltransferase	LaCOG01323	COG4468
1	2	2	1	1	1	1	1	1	1	1	1	|||||||||||	|||||||||||	567	12	14	G	GapA	Glyceraldehyde-3-phosphate dehydrogenase	LaCOG02463	COG0057
1	1	1	1	2	1	1	1	1	1	1	1	|||||||||||	|||||||||||	567	12	13	G	GlcU	Putative glucose uptake permease	LaCOG01531	COG4975
	1	1	1	1	1	1	1	1	1	1	1	-||||||||||	-||||||||||	29	11	11	G	Gnd	6-phosphogluconate dehydrogenase	LaCOG00417	COG0362
	1	1	1	1	1	1	1				1	-|||||||---|	-|||||||---|	8	8	8	G	Gnd	6-phosphogluconate dehydrogenase	LaCOG01474	COG1023





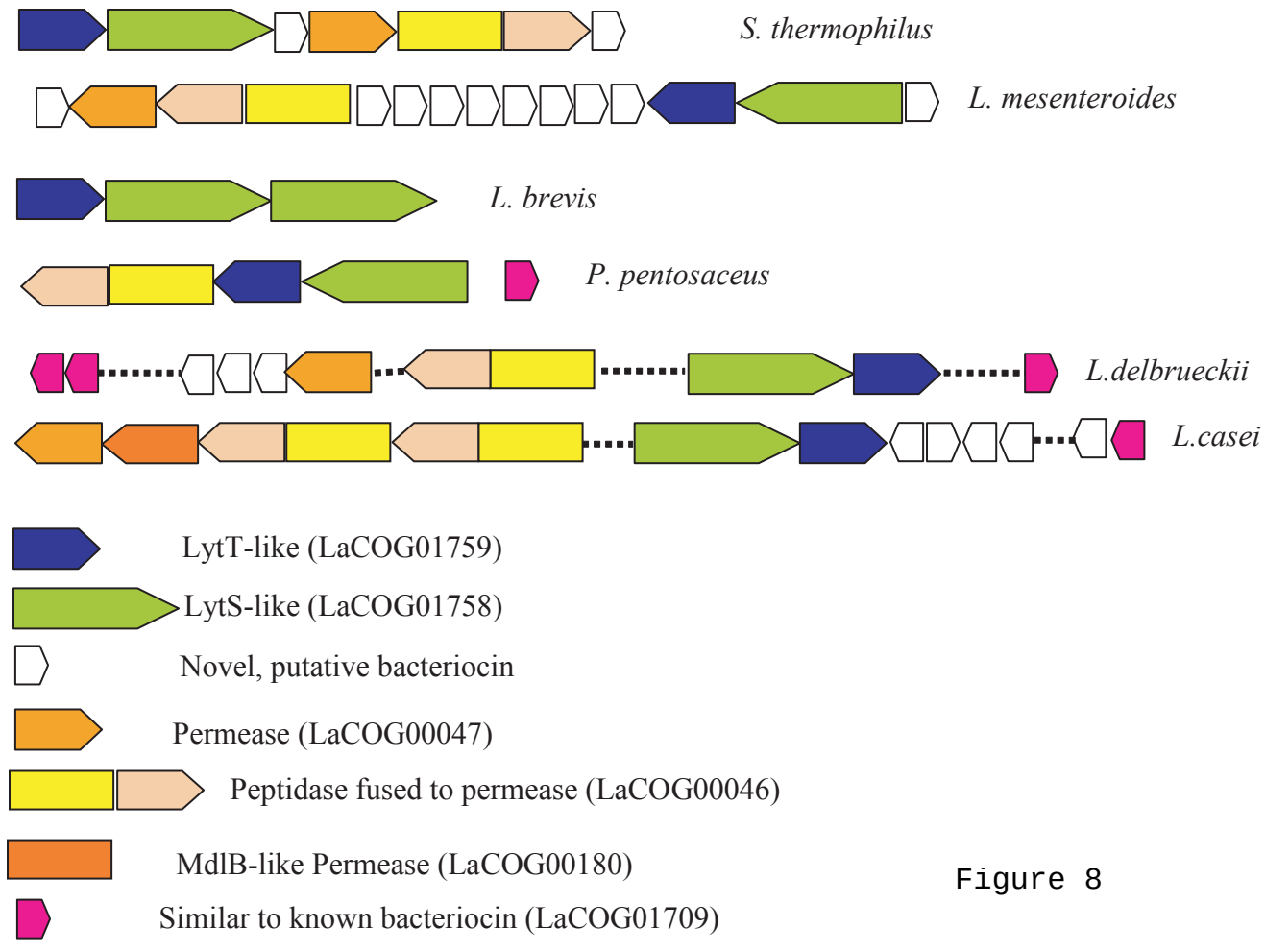

S. thermophilus
L. mesenteroides
L. brevis
P. pentosaceus
L.delbrueckii
L.casei
LytT-like (LaCOG01759)
LytS-like (LaCOG01758)
Novel, putative bacteriocin
Permease (LaCOG00047)
Peptidase fused to permease (LaCOG00046)
MdlB-like Permease (LaCOG00180)
Similar to known bacteriocin (LaCOG01709)

Figure 8

Table 8. Metabolic capabilities of *Lactobacillales*

Note: 1, present; 0, absent. Also see the same pathways assigned for LaCOGs in Table 7.

Biosynthetic pathway	***S. thermophilus***	***Lc. lactis* ssp. *lactis***	***Lc. lactis* ssp. *cremoris***	***Lb. brevis***	***Lb. plantarum***	***P. pentosaceus***	***Le. mesenteroides***	***O. oeni***	***Lb. johnsonii***	***Lb. gasseri***	***Lb. delbrueckii* ssp. *bulgaricus***	***Lb. casei***
Alanine biosynthesis	1	1	1	1	1	1	1	1	1	1	1	1
Arginine	1	1	1	0	1	0	1	0	0	0	0	0
Aromatic acids biosynthesis	1	1	1	0	1	0	1	0	0	0	0	0
Aspartate and Asparagine biosynthesis	1	1	1	?	1	1	1	1	1	1	1	1
Cysteine biosynthesis	1	1	1	0	1	0	1	1	0	0	1	1
Glutamine biosynthesis	1	1	1	1	1	1	1	1	1	1	1	1
Glutamate biosynthesis	0	1	1	0	0	0	1	0	0	0	0	1
Histidine biosynthesis	1	1	1	0	1	0	1	0	0	0	0	1
Ile, Leu, Val biosynthesis	1	1	1	0	?	0	1	?	0	0	?	?
Lysine biosynthesis	1	1	1	0	1	1	1	1	0	0	?	1
Methionine biosynthesis	1	1	1	0	1	0	1	1	0	0	1	0
Proline biosynthesis	1	1	1	0	1	0	1	0	0	0	0	0
Serine/Glycine biosynthesis	1	1	1	0	0	0	0	0	0	0	0	0
Threonine biosynthesis	1	1	1	0	1	0	1	1	0	1	1	1
Tryptophan biosynthesis	1	1	1	0	1	0	1	0	0	0	0	0
Pyrimidine biosynthesis	1	1	1	1	1	1	1	1	1	1	1	1
Purine biosynthesis	1	1	1	0	1	1	1	1	0	0	1	1
Fatty acids biosynthesis	1	1	1	1	1	1	1	1	0	0	1	1
Thiamine biosynthesis	?	1	1	?	1	1	?	0	0	0	0	1
Folate biosynthesis	1	1	1	0	1	0	0	0	0	0	1	0
Last 3 reactions of CoA	1	1	1	1	1	1	1	1	1	1	1	1
Last 3 reactions of NAD	1	1	1	1	1	1	1	1	1	1	1	1
Riboflavin biosynthesis	0	1	1	1	1	1	1	1	0	1	0	0
Thymidilate biosynthesis	1	0	0	1	1	1	1	0	0	0	0	1
Ubiquinol biosynthesis	?	1	1	0	1	0	1	?	0	0	1	0
	21	24	24	8	22	12	22	13	6	8	13	15