Lactobacillus phylogenomics – towards a reclassification of the genus

Marcus J. Claesson, Douwe van Sinderen and Paul W. O’Toole

Department of Microbiology and Alimentary Pharmabiotic Centre, University College Cork, Ireland

The extremely diverse genus Lactobacillus is the largest among the lactic acid bacteria, with over 145 recognized species. In this work, which to our knowledge is the largest comparative phylogenomics study of a single genus to date, 12 genomes of Lactobacillus strains were subjected to an array of whole-genome and single-marker phylogenetic approaches, to investigate the case for extracting subgeneric groups and to determine whether a single congruent phylogeny could be identified. We conclude that GroEL is a more robust single-gene phylogenetic marker for the genus Lactobacillus than the 16S rRNA gene, when no whole-genome information is available. Significant incongruence was found, both within a set of trees based on 141 core proteins and within those phylogenies based on numbers of orthologues, concatenated RNA polymerase subunits and single gene/protein markers. This is possibly due to different evolutionary rates, hidden paralogies or horizontal gene transfer. Such phylogenetic ambiguities are efficiently visualized with cluster-networks. Although the genus contains some highly unstable taxa, four subgeneric groups were distinguished. Qualitative and quantitative gene analysis of these groups resulted in three findings: there is a relatively small number of group-specific proteins, the majority of which are poorly characterized; major groupings are functionally better distinguishable by absent genes rather than gained/retained genes; and, finally, a gene cluster possibly involved in purine metabolism is uniquely present in four lactobacilli associated with meat. In conclusion, because of either significantly different branching patterns or the availability of too few members, three of the four identified groups could not serve as the basis for identifying candidate novel genera within the current genus. We therefore suggest targeted sequencing of key taxonomic species identified here, which are likely to add sufficient depth for a future reclassification, followed by phylogenomic analysis involving the core proteins identified here. This will ideally be combined with phenotypic data using a polyphasic approach.

INTRODUCTION

Lactobacilli are Gram-positive lactic acid bacteria (LAB) that can be found within plant, animal and foodstuff niches. Some species associated with the gastrointestinal (GI) tract are regarded as probiotic, conferring health benefits on the host. The genus Lactobacillus is the largest among the LAB, with over 145 recognized species (Euzéby, 1997). It is an extremely diverse genus, with a taxonomy that is generally regarded as unsatisfactory (Canchaya et al., 2006; Felis & DellaGio, 2007; Makarova et al., 2006). In a previous study (Canchaya et al., 2006), we compared the five genomes then available for strains of Lactobacillus. Using several methods and datasets we concluded that, although 354 core proteins were identifiable, none proved to be Lactobacillus-specific. All constructed trees highlighted the significant diversity present, and a subgeneric division was suggested, comprising three main branches. Since then, seven more genomes of members of Lactobacillus have become available. Makarova et al. (2006) reported on the significant phylogenetic and functional diversity in the Lactobacillales, mediated by extensive gene loss and horizontal gene transfer (HGT) of key genes from other LAB genomes. Hence, a revision of the taxonomy of the Lactobacillales was suggested. As partial support of these studies, Berger et al. (2007) described an extensive polyphasic approach, mainly focusing on the so-called ‘Lactobacillus acidophilus complex’, and concluded that its closest relatives were in fact Lactobacillus casei and Lactobacillus sakei.

This work is a methodological and investigative extension of the above-mentioned studies, determining the congruence produced by traditional and new phylogenomic
approaches [defined as the discipline of evolutionary analysis based on genome sequences (Eisen & Hanawalt, 1999)] of 12 sequenced lactobacilli along with six other related non-
Lactobacillus, low-G+C content, Gram-positive strains. Phylogenies were constructed using a wide array of methods, including concatenated core and RNA polymerase proteins, single-gene markers and matrices derived from orthologous gene counts. The aim was to investigate whether the resulting trees were consistent enough to propose a single unifying phylogeny and whether any stable subgeneric groups could be established. Functional analyses of these groups using clusters of orthologous groups (COGs) would then elucidate group- and possibly habitat-specific adaptations and shed further light on the characteristics of this diverse and complex genus.

METHODS

The 18 genome sequences included in this study are the publicly available sequences of Lactobacillus plantarum WCFS1 (Kleerebezem et al., 2003; GenBank accession no. AL935263), Lactobacillus johnsonii NCC533 (Pridmore et al., 2004; AE017198), L. acidophilus NCFM (Altermann et al., 2005; CP000033), L. sakei 23K (Chailolv et al., 2005; CR936503), Lactobacillus salivarius UCC 118 (Claesson et al., 2006; chromosome, CP000233; pMP118, CP000234), Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842 (van de Guchte et al., 2006; CR954253), L. delbrueckii subsp. bulgaricus ATCC BAA-365, Lactobacillus gasseri ATCC 33323 T , Lactobacillus brevis ATCC 367 and L. casei ATCC 334 (Makarova et al., 2006; CP000412, CP000413, CP000416 and CP000423, respectively), Lactobacillus reuteri F275 (unpublished; CP000705) and Lactobacillus helveticus DPC4571 (Callanan et al., 2008; CP000517). In addition, the genome sequences of Pedococcus pentosaceus ATCC 25745, Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293 and Oenococcus oeni PSU-1 (Makarova et al., 2006; GenBank accession numbers CP000422, CP000414 and CP000411, respectively) were added because of their well-recognized close similarity to the genus Lactobacillus, and the sequences of Streptococcus thermophilus CNRZ1066 (Botolin et al., 2004; CP000024), Bacillus subtilis 168 (Kunst et al., 1997; AL009126) and Enterococcus faecalis V583 (Paulsen et al., 2003; AE016830) were included as outgroups.

Four criteria were used for identifying core proteins: (i) sharing the same putative orthologues [as defined by reciprocal best BLAST hits (Pearson & Lipman, 1988) with a minimum of 30% identity over at least 80% of both protein lengths] with all of the other genomes; (ii) belonging to a protein family containing exactly 18 genomes as clustered by the graph-theory-based MCL algorithm (Enright et al., 2002) using a BLAST (Altschul et al., 1990) cut-off E-value of \( 1 \times 10^{-16} \); (iii) all members of a core protein must have the same COG category assigned; and (iv) obvious mobile elements, such as transposases, plasmid- and phage-related proteins, were removed. Sequences were then aligned with MUSCLE (Edgar, 2004) where gapped regions had been removed. All phylogenetic trees, except the neighbour-joining tree based on numbers of orthologues, were built by the maximum-likelihood method using PhyML (Guindon & Gascuel, 2003) with 100 bootstrap runs. Appropriate substitution models were selected by modelGenerator (Keane et al., 2006) (GTR for nucleotide and WAG for amino acid sequences). All trees were manually outgrouped with B. subtilis. A neighbour-joining tree based on gene content was constructed from distances \( D_{AB} \), which is a function of the number of orthologues, normalized by the weighted mean of the total number of proteins (excluding paralogues and obvious mobile elements), between each of the 153 possible genome pairs (Korbel et al., 2002):

\[
D_{AB} = -\ln \left( \frac{\sum_{i=1}^{n} (all\ genes)_A + (all\ genes)_B}{2 \times (all\ genes)_A \times (all\ genes)_B} \right) \times (shared\ orthologues)_{AB}
\]

Finally, we used SplitsTree (Huson & Bryant, 2006) and MEGA (Tamura et al., 2007) to visualize the phylogenetic trees and cluster- and supernetworks. Robinson–Foulds distances were calculated using TreeDist from the PHYLIP package (Retief, 2000). By implementing various Perl scripts followed by manual inspection of BLAST searches against the nr protein database, group- and habitat-specific protein gains/retentions or absences could be identified. Most of the included genomes had been annotated with COGs through the NCBI website (Tatusov et al., 2003), but for those with missing COGs, the following criteria were used to obtain them: (i) each proteome was searched against the COG database with a cut-off E-value of 0.01, which renders a number of possible COGs for each protein; (ii) a COG assignment was accepted automatically if the top BLAST hit was the same as the majority of the first 20 hits; (iii) finally, for all cases where the second test failed, or where only a single hit was detected, COG assignments were curated manually and adjusted in accordance with the existing annotation. When additional functional evidence was needed, proteins were searched with InterProScan (Quevillon et al., 2005) against databases with known domains, motifs or other functional signatures.

RESULTS AND DISCUSSION

Phylogenetic analysis

In order to produce robust trees and to minimize phylogenetic incongruence, it is essential to select carefully core proteins that are ubiquitous across the genomes in question and on which the phylogenies will be based. The simple inclusion of as many genes as possible can result in tree-reconstruction artefacts and incongruence between phylogenies: these misleading effects may be due to duplicated or horizontally transferred genes or to systematic errors resulting from inference methods or non-phylogenetic signals, such as nucleotide compositional bias (Jeffroy et al., 2006; Philippe & Douady, 2003). Hence, to minimize these errors, a very careful identification of core proteins was done, using four separate criteria that are described in Methods. Apart from applying the strict rule of only accepting the intersection of two separate sequence-based orthology methods, this regimen ensured that no paralogues were present among the core proteins. Protein sequences were used preferentially rather than nucleotide sequences, because of their better conservation across such a diverse genus. Accordingly, 141 core proteins were identified, from which a tree was constructed using the alignments of their concatenated and untagged protein sequences encompassing 46 465 positions (Fig. 1a). All core proteins except one, methionyl-tRNA synthetase (COG143), were present among the 567 LaCOGs previously described by Makarova et al. (2006) from 12 analysed genomes. The significant difference in numbers [141 here compared with 567 in the paper of Makarova et al. (2006)] can be explained by the non-overlapping genome sets, where we additionally included L. reuteri, L.
Fig. 1. (a–e) Phylogenetic trees based on sequences of 141 concatenated core proteins (a), four RNA polymerase subunits (b), 16S rRNA genes (c) and GroEL proteins (d) and number of orthologues (e). Branches whose short lengths are a possible cause of unstable taxa are marked with arrows. (f) Cluster-network based on a combination of 141 separate gene trees. All possible splits are indicated as extra edges. *L. bulgaricus* 1 corresponds to strain ATCC 11842T and *L. bulgaricus* 2 to strain ATCC BAA-365. *Leu., Leuconostoc*. 

http://ijs.sgmjournals.org 2947
and another L. bulgaricus strain. Moreover, our criteria for defining a core protein are also stricter than those that were used for assigning a protein to a certain LaCOG. As a consequence, a smaller set of more carefully identified core proteins ensures a smaller number of false positives, which is important for the subsequent phylogenetic analysis.

The phylogenies constructed by Makarova et al. (2006) showed a division of Lactobacillus into two distant groups: L. brevis, L. plantarum and P. pentosaceus in the first group and a second including L. bulgaricus, L. gasseri and L. johnsonii, to which L. casei is basal. An additional branch with Leuconostoc mesenteroides and O. oeni was located between the two clades. In a subsequent paper, these trees were complemented by adding L. acidophilus, L. sakei and L. salivarius along with four streptococci (Makarova & Koonin, 2007). The resulting phylogeny had the same topology as in their previous trees, except that L. salivarius was outside and basal to the Leuconostoc mesenteroides/O. oeni pair.

For a comparison to these studies, we built a tree from the four concatenated RNA polymerase subunits (Fig. 1b), which are a subset of the 141 core proteins (2797 positions). This tree showed only one slight topological difference from the concatenated RNA polymerase tree from the Makarova study; in our analysis, L. brevis is grouped with L. plantarum, whereas it is immediately outside L. plantarum and P. pentosaceus in the Makarova tree. This is probably due to differences in taxon size and/or tree-building methodologies. In addition, there are two noteworthy differences in topology between our trees based on the 141 core proteins and the RNA polymerase proteins: the L. bulgaricus pair in the ‘core protein’ tree is outside the other species in the ‘acidophilus complex’, while being internal in the RNA polymerase tree; and the O. oeni/Leuconostoc mesenteroides pair is basal to the L. salivarius subgroup in the core protein tree, rather than being deeply internal in the RNA polymerase tree. Since the latter clade displays poorer bootstrap support, its internal topology may be less trustworthy, while the core protein tree appears to afford greater resolution. Together with the fact that the core protein tree is based on a larger number of carefully selected proteins, we therefore propose that it is more suitable for inferring Lactobacillus phylogeny than the RNA polymerase tree.

It should be mentioned, however, that the generally high bootstrap values, which are indicative of a strongly supported tree, do not necessarily mean that these trees are ‘correct’; the long lengths of the concatenated sequences, which are primarily responsible for the strong support, merely reduce stochastic and sampling errors common in traditional single-gene phylogenies, while having little impact on the more cumbersome systematic errors (Delsuc et al., 2005; Holland et al., 2004; Jeffroy et al., 2006).

Another sequence-based phylogenomic inference method is based on combining individual gene trees, each of which can also contain non-overlapping sets of taxa, into a supertree (Bininda-Emonds, 2004). The 141 trees in our dataset proved to be far too disparate to produce an optimal supertree, according to results produced by the Clann software (Creevey & McInerney, 2005) (data not shown). Instead, we used a so-called cluster-network approach, which is a recently developed type of phylogenetic network for representing non-hierarchical cluster systems (Daniel Huson, personal communication). Thus, Fig. 1(f) shows a rooted network where each split between taxa is represented by a single edge. This visualization reveals inconsistencies between single-gene topologies, thereby highlighting clades that are more incongruent than others. Another way of simultaneously displaying contradictory evolutionary relationships is by means of a rooted supernetwork (Huson et al., 2004), which we also applied to the same dataset (Supplementary Fig. S1, available in IJSEM Online). This type of network shows the same incongruence, but represents incompatible splits as parallelograms in the graph. Possible reasons behind these conflicting topologies could be differences in evolutionary rates and/or the occurrence of HGT (Philippe & Douady, 2003). However, although still a debated area, it has recently been proposed that HGT has limited impact on whole-genome phylogenies (Choi & Kim, 2007; Eisen & Fraser, 2003). When also considering the fact that obvious HGT elements were excluded from our core protein set, we find it more likely that other systematic errors are involved, such as differential evolutionary rates and possibly hidden paralogies. If, on the other hand, genetic exchange through HGT ultimately proves to be a sufficiently strong evolutionary force to influence bacterial phylogeny significantly, phylogenetic networks would be more suitable to represent bacterial evolutionary lineages properly (Dagan & Martin, 2006; Huson & Bryant, 2006). Hence, this would further support the need for alternative phylogenetic models to the traditional trees, for which congruence is a necessity.

Finally, Robinson–Foulds distances (Robinson & Foulds, 1981) between the generated trees were calculated as a quantitative measurement of the topological differences (Supplementary Table S1). These numbers of bipartitions, which are different between trees, thereby confirm the observed significant incongruence.

The phylogenetic incongruence of several of the taxa studied suggests an unclear evolutionary path for many Lactobacillus lineages. Nevertheless, based on these phylogenies, it is still possible to distinguish the following subgeneric groups:

- **Group A** (L. acidophilus, L. helveticus, L. bulgaricus, L. johnsonii and L. gasseri)
- **Group B** (L. salivarius, L. plantarum, L. reuteri, L. brevis and P. pentosaceus)
- **Group C** (L. sakei and L. casei)
- **Group D** (Leuconostoc mesenteroides and O. oeni)

The main question here is whether these subgroups represent phylogenetic entities reliable enough to serve as
a basis for a reclassification of the genus *Lactobacillus*. Compared to the topological differences within the groups (A and B), there is relatively little variation between them; group C is basal to group A and group D is basal, or almost basal, to group B. There are, however, a number of problems with such a genus reclassification at present. Group B displays significantly different internal branching patterns across the generated phylogenies and is not as well-defined as group A. In addition, groups C and D are populated by only two species each and are therefore not representative enough to form the basis for new genera. Last but not least, there are no clear ecological niche associations, which can be exemplified best by the fact that *L. acidophilus* is found within the GI tract, and its closest relative *L. helveticus* in cheese. Similarly, *L. sakei* is associated with fermented meat products whereas *L. casei* is found in dairy products, *L. salivarius* in the GI tract and *P. pentosaceus* on meat and plants. This highlights an important and inherent problem with correlating phylogenomic and phylogenetic relatedness with physiological, morphological and environmental similarities. The former phylogenies are based on genes that are ubiquitous throughout the taxa, whereas phenotypic differences usually arise from differences in gene content. Hence, polyphasic approaches (integrating phenotypic, genotypic and phylogenetic data) to consolidate an unsatisfactory taxonomy like this represent a tremendous challenge, but are nevertheless necessary for a complete and final resolution. Furthermore, to overcome the phylogenetic difficulties mentioned above, more species need to be included at key taxonomic positions. At present, genomes of four strains of *L. reuteri*, three strains of *Lactobacillus rhamnosus*, two strains of *L. casei*, *Lactobacillus crispatus* and *L. helveticus* and one strain each of *Lactobacillus amylophilicus*, *L. antri*, *L. bulgaricus*, *L. buchneri*, *L. coleohominis*, *L. fermentum*, *L. iners*, *L. jensenii*, *L. paracasei*, *L. sakei*, *L. sanitoryeus*, *L. ultunensis*, *Leuconostoc mesenteroides* and *O. oeni* are either completely sequenced but not yet published or are being sequenced according to the Genomes OnLine Database (Liolios et al., 2008). Apart from the additional strains of the already represented species *L. reuteri*, *L. casei*, *L. sakei*, *Leuconostoc mesenteroides* and *O. oeni*, of which the latter four will further populate the dual-inhabitant groups C and D, *L. antri*, *L. buchneri*, *L. coleohominis* and *L. fermentum* are likely to be welcome additions to group B and *L. rhamnosus* and *L. paracasei* to group C [predicted from previously published 16S rRNA gene phylogenies (Canchaya et al., 2006; Felis & Delliglao, 2007)]. Furthermore, certain carefully selected species (preferably represented by type strains) could prove to be of extra significance for a future phylogenetic consolidation, such as those that are likely to be positioned between groups or which represent taxonomic clades without any sequenced members according to the 16S rRNA gene-based phylogenies. Some examples of such species are *Lactobacillus perolens*, *L. animalis*, *L. algidus*, *L. lindneri*, *L. amylophilus* and *L. alimentarius*. Species within the other closely related genera *Leuconostoc*, *Oenococcus*, *Pediococcus* and *Paralactobacillus* are also important targets for genome sequencing and phylogenomic analysis, in order to exclude or include these genera finally from a future reclassification. Encouragingly, the rapid increase in speed and decrease in cost of *de novo* bacterial genome sequencing over recent years offer opportunities, better than ever before, to sequence strains of a sufficient number of phylogenetically important species related to *Lactobacillus*, whose genomes could provide enough data for a reclassification. In fact, even if a genome sequence is not completely finished and circularized, which is the case especially for many commercially sequenced strains, sequences of the vast majority of core proteins can be extracted and compared with the set of core proteins presented here. Phylogenomic analysis of these additional genomes may then finally answer the important question whether *Lactobacillus* subgroups form natural clusters or are just a genetic continuum of diversity that is refractile to fragmented grouping.

As a further demonstration of the extensive diversity within the genus, we looked at a ‘traditional’ way to distinguish between species and genera, 16S rRNA gene sequence similarity, which applies as a rule of thumb that organisms share >97 % of this sequence at a species level and >95 % at a genus level (Rossello–Mora & Amann, 2001). Remarkably, the only organisms that fulfill either criterion were *L. acidophilus*–*L. helveticus* (98.4 %), *L. johnsonii*–*L. gasseri* (99.6 %) and the two *L. bulgaricus* strains (99.7 %). These are subsequently followed by 16S rRNA gene sequence similarities between *L. plantarum*–*L. brevis* (94.2 %), *L. plantarum*–*L. sakei* (93.4 %) and *L. acidophilus*–*L. bulgaricus* (93.7 %). Thus, all investigated species not belonging to group A are outside the boundaries for genus delineation. However, it has to be pointed out that the prokaryotic species concept can be seen as both problematic and arbitrary, and is the subject of ongoing debate (Gevers et al., 2005; Konstantinidis et al., 2006). Another method that is used for species delineation is DNA–DNA hybridization, using 70 % relatedness as a threshold. However, owing to technical difficulties, this method may be replaced in future by the simpler and more accurate technique of comparing average nucleotide or amino acid identities (ANI/AAI) between orthologues in prokaryotic species. As a further demonstration of the extensive diversity within the genus, we looked at a ‘traditional’ way to distinguish between species and genera, 16S rRNA gene sequence similarity, which applies as a rule of thumb that organisms share >97 % of this sequence at a species level and >95 % at a genus level (Rossello–Mora & Amann, 2001). Remarkably, the only organisms that fulfill either criterion were *L. acidophilus*–*L. helveticus* (98.4 %), *L. johnsonii*–*L. gasseri* (99.6 %) and the two *L. bulgaricus* strains (99.7 %). These are subsequently followed by 16S rRNA gene sequence similarities between *L. plantarum*–*L. brevis* (94.2 %), *L. plantarum*–*L. sakei* (93.4 %) and *L. acidophilus*–*L. bulgaricus* (93.7 %). Thus, all investigated species not belonging to group A are outside the boundaries for genus delineation. However, it has to be pointed out that the prokaryotic species concept can be seen as both problematic and arbitrary, and is the subject of ongoing debate (Gevers et al., 2005; Konstantinidis et al., 2006). Another method that is used for species delineation is DNA–DNA hybridization, using 70 % relatedness as a threshold. However, owing to technical difficulties, this method may be replaced in future by the simpler and more accurate technique of comparing average nucleotide or amino acid identities (ANI/AAI) between orthologues in separate genomes (Konstantinidis et al., 2006). A species cut-off value of 98 % ANI will then correspond to 70 % DNA–DNA relatedness, and thus the value serves as a suitable species definition (Goris et al., 2007). These criteria, however, cover only the species concept, and unfortunately do not provide guidance on how to delineate between different genera, which would be more significant for the present study.

Phylogenetic relationships between the more than 100 recognized *Lactobacillus* species have traditionally been inferred using alignments of 16S rRNA gene sequences (Canchaya et al., 2006; Delliglao & Felis, 2005; Felis & Delliglao, 2007), which is by far the most common single-
gene phylogenetic marker employed for prokaryotes. These phylogenies have not resolved the taxonomic issues, however, since they often conflict with the older, phenotypic classification (Vandamme et al., 1996). Here, we used the same methodology as in our previous work, extending our analysis to the 18 species included in this study. Since the number of rRNA operons varies from two in O. oeni to ten in B. subtilis, we used the consensus nucleotide sequence for each taxon, rendering an ungapped alignment of 1298 positions (Supplementary Table S2 shows the mean percentage identities between 16S rRNA genes). Surprisingly, the resulting topology shown in Fig. 1(c) deviates from the aforementioned subgeneric groupings, in that group C (L. sakei and L. casei) is wedged between L. salivarius and the rest of group B. L. reuteri is also immediately outside group A and C instead of being grouped with group B and D. Much of the internal topology has short branch lengths (as indicated with arrows in Fig. 1c) and remarkably low bootstrap support, which makes this tree highly unstable. Similar conclusions were drawn from our previous studies (Canchaya et al., 2006, data not shown) and those of Felis & Dallaglio (2007): when inserting additional 16S rRNA genes into the species tree, some of the internal poorly supported topologies, and especially the position of L. sakei, would change dramatically. Another observation of instability was also made when comparing the 16S rRNA gene consensus tree with the tree containing all 16S rRNA genes (data not shown). Reassuringly, however, with the exception of the two L. bulgaricus strains, all 16S rRNA genes grouped together within their respective species nodes. The 14 groups established by Felis & Dallaglio (2007), based on a 16S rRNA gene tree including eight Pediococcus strains, are not easily compared with our 16S rRNA gene tree only covering the four subgroups which, in addition, contain Leuconostoc and Oenococcus strains, due to their finer granularity. It is clear, however, that our group A is peripheral in both trees and corresponds to the L. delbrueckii group in the phylogeny of Felis & Dallaglio (2007).

Because of these shortcomings in the 16S rRNA gene phylogeny, we investigated an alternative phylogenetic marker, the well-conserved and ubiquitous chaperonin GroEL, which is considered a valuable alternative to the 16S rRNA gene for phylogenetic reconstructions in bacteria (Ludwig & Schleifer, 1999). Previous studies, such as in bifidobacteria (Jian et al., 2001; Ventura et al., 2004) and lactobacilli (Bringel et al., 2005), have showed its usefulness in determining and discriminating between species. Although highly conserved, GroEL was shown to provide better phylogenetic resolution than the 16S rRNA gene because of its greater sequence variation between related bacteria. This also proved to be the case in the present study, where the GroEL alignment of 533 amino acids rendered a tree (Fig. 1d) that had longer internal branches than the 16S rRNA gene tree, as well as higher bootstrap values, and was in better agreement with the concatenated sequence trees. The two main differences from the 16S rRNA gene phylogeny are that L. casei is immediately outside and not grouped with L. sakei in the GroEL tree and that L. plantarum has a more peripheral position close to L. salivarius in the GroEL tree. Accordingly, we would therefore suggest that the GroEL protein is a more robust and suitable phylogenetic marker than the 16S rRNA gene for unsequenced members of the genus Lactobacillus.

Finally, to explore a tree-building method that was not dependent on multiple sequence alignments, we constructed a neighbour-joining tree based on orthologous gene counts, which were normalized by the weighted mean of the total number of proteins for each of the 153 possible genome pairs (Fig. 1e). The calculated distances and the numbers of orthologues detected, with and without paralogues and mobile elements removed, are shown in Supplementary Table S3. Clearly, our proposed grouping scheme remains intact, but with some differences between groups; L. sakei and L. casei are no longer basal to group A and group D is external to group B, as well as the other two groups. Although this methodology does not provide statistical support, it is noteworthy that the short lengths, indicated by arrows in the tree, could be responsible for these intergroup deviations from our previously generated trees, and that the topology in fact could be very similar to the one in Fig. 1(a), based on 141 core proteins.

Conservation of gene order (synteny) is another whole-genome feature that phylogenies can be reconstructed from. However, due to the extensive diversity of the genus, such an analysis would not be very meaningful here.

**Variably present Lactobacillus proteins**

COG assignments were either annotated by us, for half of the genomes, or obtained from the NCBI website. These allowed protein functions to be compared and analysed in a structured way across the whole genome set, at levels of both fine and coarse detail. The COG database contains a relatively small number (66) of genomes, compared with the total number of genomes available. However, these COGs contain conserved functional protein domains from model organisms that have been widely used in numerous genome projects. Should many more genomes be included in the COG database, it is likely that these would only increase COG descriptions such as ‘unknown function’ for proteins that are now without COG assignments, and hence be of little use in this study. Table 1 is a summary of the top-level COG categories for a number of protein collections, and Supplementary Table S4 shows a breakdown of the corresponding locus tags into functional annotation and low-level COGs. Not surprisingly, a majority of the 141 core proteins with assigned function have housekeeping roles, although a relatively large proportion are also involved in post-translational modification, nucleotide metabolism and transport, as well as cell-wall biogenesis.
There was a relatively small number of proteins specific to any of the four proposed subgeneric groups (Table 1); groups A and D have the largest numbers of group-specific proteins, 6 and 10, respectively, as would be expected from their relative stability and longer evolutionary distances (branch lengths) from the two other groups. However, groups A, B and C have only poorly characterized proteins that are group-specific, while group D also contains proteins involved in transcription, translation, cell-wall biogenesis and defence mechanisms. Some of the more interesting proteins that potentially add to the understanding of the evolution of the different groups are as follows: for group A, a metallophosphoesterase, an abortive infection protein and a predicted kinase; for group C, a hypothetical cell-surface protein precursor; and for group D, a sugar phosphate isomerase involved in capsule formation, a 3-deoxy-D-manno-octulosonic acid (KDO) 8-phosphate synthase related to proteins involved in lipopolysaccharide biosynthesis, a peptidoglycan interpeptide bridge-formation enzyme and two predicted membrane proteins. It is tempting to suggest that these cell-envelope-related proteins would help to explain morphological differences between these two species and the other lactobacilli.

In contrast, there seems to be a much higher degree of gene absence between the groups, as was also highlighted previously (Makarova et al., 2006). Because of the unclear evolutionary history of these groups, we can only define genes absent from a certain group, but present in the other three groups, as group-specific absent genes, rather than group-specific lost genes. Accordingly, 18 genes were absent from group A and 17 from group D, but none from groups B and C. This is probably due to the relative stability and sustained evolution of groups A and D, also reported by Berger et al. (2007). The functional categories of proteins associated with absent genes are more widely spread than those of group-specific proteins, as can be seen in Table 1. Some of the more interesting absences are: from group A, a pyruvate/2-oxoglutarate dehydrogenase complex, lipoate-protein ligase A, universal stress protein UspA, lysophospholipase L1 involved in glycerophospholipid metabolism, a diglycerol kinase, a queuine/archaeosine tRNA-ribosyltransferase, a glutaredoxin and a cytochrome c; and from group B, an asparagine synthetase, a metal-dependent hydrolase of the beta-lactamase superfamily, I, a lipoprotein signal peptidase, thiamine biosynthesis ATP pyrophosphatase and a prolineproline diacylglyceroltransferase. More detailed information can be found in Supplementary Table S4.

Because of the different numbers of species analysed, an exact comparison is not possible with the previous study of gene gain/retention and loss by Makarova et al. (2006). However, the majority of the 21 gene losses in group A and 18 gene losses in group D reported by Makarova et al. (2006) correspond with our findings (not shown).

Table 1. Summary of top-level COG categories for a number of protein collections

The total number of core protein COGs is 144 and not 141 because three proteins each have two COGs assigned. Further details are given in Supplementary Table S4.

<table>
<thead>
<tr>
<th>COG category</th>
<th>Core proteins</th>
<th>Group-specific proteins</th>
<th>Group-specific absence</th>
<th>Lactobacillus-specific proteins</th>
<th>Meat-specific proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>A  B  C  D</td>
<td>A  D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Translation</td>
<td>59</td>
<td>1 1 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcription</td>
<td>9</td>
<td>1 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replication, recombination and repair</td>
<td>15</td>
<td>1 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell cycle control, mitosis and meiosis</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Defence mechanisms</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Signal transduction mechanisms</td>
<td>3</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell-wall/membrane biogenesis</td>
<td>9</td>
<td>2 1 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intracellular trafficking and secretion</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-translational modification</td>
<td>10</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Energy production and conversion</td>
<td>6</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carbohydrate transport and metabolism</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amino acid transport and metabolism</td>
<td></td>
<td>2 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nucleotide transport and metabolism</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coenzyme transport and metabolism</td>
<td>7</td>
<td></td>
<td>1 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lipid transport and metabolism</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secondary metabolite biosynthesis</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General function prediction only</td>
<td>9 1</td>
<td>1 3</td>
<td>2 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Function unknown</td>
<td>1 1 4 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No COGs</td>
<td>5 2 3 4 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>144 6 2 10 18</td>
<td></td>
<td>18 17 4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Altogether, these findings can be seen as further support for the hypothesis that gene loss, in contrast to the less-frequently observed gene gain, has shaped the evolutionary landscape of bacteria in general (Ochman, 2005) and of the genus Lactobacillus in particular (Makarova et al., 2006).

Since so few proteins seem to be unique within these phylogenetically close groups, we decided to investigate the existence of potentially habitat-specific proteins, which could be found across groups but were present within certain environmental niches. These possible habitats are the GI tract (traditionally associated with L. acidophilus, L. gasseri, L. johnsonii, L. salivarius and L. reuteri), plants (L. plantarum, L. brevis, P. pentosaceus and Leuconostoc mesenteroides), meat (L. plantarum, L. brevis, P. pentosaceus and L. sakei) and dairy products (L. helveticus, L. bulgaricus and L. casei). Only four habitat-specific proteins, unique to Lactobacillus species found on meat, were found: a PurE-related NCAIR mutase possibly involved in purine biosynthesis, an ATP-utilizing enzyme, a hypothetical protein and a transcription regulator belonging to the Crp family. The first three proteins were all conserved within a gene cluster across the four genomes, which also contains genes for a glycerol uptake facilitator protein, another hypothetical protein and a transcription-regulator belonging to the Crp family. A similar gene cluster with a slightly different gene order, and without the Crp regulator, can also be found in Staphylococcus epidermis, Staphylococcus saprophyticus (skin pathogens), Wolinella succinogenes (non-pathogen in bovine and human gut) and Corynebacterium efficiens (soil). In L. sakei, this cluster is less than 4 kb upstream of a purine-nucleoside phosphorylase along with two nucleoside and purine transporters (Supplementary Table S2 and Chaillou et al., 2005). These findings point towards an essential locus, potentially necessary for scavenging nucleotide precursors in these lactobacilli.

Finally, again underlining the extreme diversity of the genus, we found only one potential Lactobacillus-specific protein: the conserved hypothetical cytosolic protein (LSL_1300 in L. salivarius) was the only protein present throughout all four groups and without any other homologues with E-values less than 1. A weak hit against the superfamily SSF54928 is an indication that it could have an RNA-binding function.

**Concluding remarks**

Phylogenetic trees based on single genes/proteins are useful for rapid positioning of large groups of taxa with unsequenced genomes. Because of dissimilarities of the (poorly supported) tree based on 16S rRNA gene sequences from those constructed from multiple genes and the relative robustness of the GroEL tree, we suggest that GroEL is a more suitable phylogenetic single-gene marker than the 16S rRNA gene for large-scale analysis.

For focused analysis of taxa of suspected proximity, phylogenies based on whole genomes may be more discriminatory and ultimately more informative. To summarize the multigene phylogenies generated in this study, there are significant topological differences between the generated trees. This is also visualized clearly in the cluster- and supernetworks combining phylogenies from 141 core proteins. The source of incongruence is so far unknown, but it is likely that incompatible inference methods, different evolutionary rates among the genes, deleted paralogues and/or, to a smaller extent, HGT play important roles. It is therefore possible that the genus Lactobacillus contains some persistent highly unstable taxa which are perhaps better represented by network structures.

The extreme diversity across the genus Lactobacillus was highlighted in this study and, since most lactobacilli were in breach of the traditional genus definition according to 16S rRNA gene sequence similarities, and only one Lactobacillus-specific protein was found, it is necessary to investigate the feasibility of a reclassification of the genus into new genera, based on the phylogenetic data available at this point. Of the four identified subgeneric groups, only that corresponding to the ‘L. acidophilus complex’ (group A) is large and well-defined enough to distinguish it clearly from the others. However, sequence data from certain ongoing genome projects, as well as targeted sequencing of key taxonomic species identified here, are likely to add sufficient depth for a future phylogenetic consolidation. In addition to this, more polyphasic approaches are needed, which combine phylogenetic data with phenotypic characteristics. Therefore, we call for collaborations between laboratories working on taxonomies related to the genus Lactobacillus to make this possible. Gene absence, much more than gene gain, seems to have affected the evolution of these groupings, especially for groups A and D. The only loci found exclusively in species associated with a certain habitat were four proteins possibly involved in purine metabolism; these were found across a cluster in L. plantarum, L. brevis, P. pentosaceus and L. sakei, which all are capable of surviving on meat. Closer analysis of these and other group-related proteins will give further clues into the evolutionary history and environmental adaptations that have shaped the members of this fascinating and complex genus.

**ACKNOWLEDGEMENTS**

This research was supported by Science Foundation Ireland through a Centre for Science, Engineering and Technology award to the Alimentary Pharmabiotic Centre. We thank Peter J. Lockhart, Daniel Huson and Carlos Canchaya for advice on various phylogenetic methodologies. M.J.C. and P.W.O. designed the research, M.J.C. collected the data and performed the analysis and M.J.C. wrote the draft manuscript, which was revised by P.W.O and D. vanS.

**REFERENCES**

Altermann, E., Russell, W. M., Azzacate-Peril, M. A., Barrangou, R., Buck, B. L., McAuliffe, O., Souther, N., Dobson, A., Duong, T. & other authors (2005). Complete genome sequence of the probiotic lacti
acid bacterium *Lactobacillus acidophilus* NCFM. *Proc Natl Acad Sci U S A* 102, 3906–3912.


