Diversity of the genus *Lactobacillus* revealed by comparative genomics of five species

Carlos Canchaya,‡ Marcus J. Claesson,‡ Gerald F. Fitzgerald, Douwe van Sinderen and Paul W. O’Toole

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

The genus *Lactobacillus* contains over 80 recognized species, and is characterized by a high level of diversity, reflected in its complex phylogeny. The authors’ recent determination of the genome sequence of *Lactobacillus salivarius* means that five complete genomes of *Lactobacillus* species are available for comparative genomics: *L. salivarius*, *L. plantarum*, *L. acidophilus*, *L. johnsonii* and *L. sakei*. This paper now shows that there is no extensive synteny of the genome sequences of these five lactobacilli. Phylogeny based on whole-genome alignments suggested that *L. salivarius* was closer to *L. plantarum* than to *L. sakei*, which was closest to *Enterococcus faecalis*, in contrast to 16S rRNA gene relatedness. A total of 593 orthologues common to all five species were identified. Species relatedness based on this protein set was largely concordant with genome synteny-based relatedness. A *Lactobacillus* supertree, combining individual phylogenetic trees from each of 354 core proteins, had four main branches, comprising *L. salivarius–L. plantarum; L. sakei; E. faecalis*; and *L. acidophilus–L. johnsonii*. The extreme divergence of the *Lactobacillus* genomes analysed supports the recognition of new subgeneric divisions.

INTRODUCTION

The genus *Lactobacillus* is the largest group among the *Lactobacteriaceae*, and contains over 100 species (Dellaglio *et al.*, 2005; Satokari *et al.*, 2003). Lactobacilli are members of the lactic acid bacteria, whose primary fermentation end product is lactic acid (Tannock, 2004). Lactobacilli are nutritionally fastidious, and are associated with a large variety of plants and animals, a factor that has presumably contributed to their diversity, by adaptive radiation. Lactobacilli are used extensively for fermentation of plant material, dairy products and meat (reviewed by Stiles, 1996), and thus have been intensively investigated for their industrial applications (Konings *et al.*, 2000). In addition, some species have potential for production of raw ingredients for industrial processes, such as propanediol production for the textile industry (Nakamura & Whited, 2003). Lactobacilli are part of the normal human gastrointestinal (GI) microbiota and they may also be found in the GI tracts of other mammalian species (Klaenhammer & Russell, 2000; Tannock, 2004; Vaughan *et al.*, 2005). Some *Lactobacillus* species have been attributed with probiotic properties, implying ‘living micro-organisms which upon ingestion in certain numbers exert health benefits beyond inherent nutrition’ (Guarner & Schaafsma, 1998). This has added further incentive to detailed microbiological, biochemical and genomic studies of lactobacilli (Klaenhammer *et al.*, 2005). Taxonomic analysis has already led to a recognition of the unusual diversity of the genus *Lactobacillus* (Dellaglio *et al.*, 2005; Dellaglio & Felis, 2005b), and one objective of this study was to extend the 16S rRNA phylogeny of the lactobacilli (Dellaglio & Felis, 2005) and investigate its correlation with genome-based comparison.

Beginning with *Lactobacillus plantarum* WCFS1 in 2003 (Kleerebezem *et al.*, 2003), five *Lactobacillus* genome sequences are now available. *L. plantarum* WCFS1 was isolated from human saliva, and is a metabolically diverse organism with a relatively large genome of 3.3 Mb (Kleerebezem *et al.*, 2003). This species is used for plant fermentation to produce silage (Cai *et al.*, 1999), but also has probiotic properties (Molin, 2001), and its metabolic diversity was explained at a molecular level by the complexity of its genome (Kleerebezem *et al.*, 2003). *Lactobacillus johnsonii* NCC533 was isolated from human faeces, and has been thoroughly characterized for probiotic properties (reviewed by Pridmore *et al.*, 2004). The 1.99 Mb genome sequence of *L. johnsonii* NCC533 (Pridmore *et al.*, 2004) highlighted the host dependence of this multiply auxotrophic strain and its potential for host interaction via surface proteins. The *Lactobacillus acidophilus* NCFM genome sequence of 1.99 Mb (Altermann *et al.*, 2005) revealed the genetic basis for the biochemical complexity and host-interaction capability of this probiotic strain,

‡These authors contributed equally to this work.

Abbreviation: GI, gastrointestinal.
which has been employed extensively by the dairy industry. The recent genome sequence determination for *Lactobacillus sakei* strain 23K (Chaillou et al., 2005) provided a contrasting insight into the adaptive mechanisms of this meat-associated organism with extensive auxotrophy, including purine nucleoside scavenging, resilience against changing redox and oxygen levels, and potential mechanisms for biofilm formation.

We recently sequenced the genome of *Lactobacillus salivarius* UCC118 (Claesson et al., 2006). This strain was isolated from the terminal ileum of a healthy human subject undergoing reconstructive urinary tract surgery (Dunne et al., 2002; Molin et al., 1993; Rogosa et al., 1953). The 2·13 Mb genome of *L. salivarius* UCC118 comprised a 1·83 Mb chromosome, a 242 kb megaplasmid and two smaller plasmids of 20 kb and 44 kb. Annotation of the genome sequence indicated an unexpected level of prototrophy compared to other enteric lactobacilli, partly due to the contribution of megaplasmid-encoded genes, including those for completion of the pentose-phosphate pathway (Claesson et al., 2006). The availability of the genome sequence is enabling a systems approach to investigate the interaction of *L. salivarius* UCC118 with the human GI tract, exemplified by genome-wide analysis of sortase-anchored adhesins (van Pijkeren et al., 2006).

Preliminary comparative analysis of the total genome of this strain to other *Lactobacillus* genomes suggested a lack of extensive genome conservation (Claesson et al., 2006). Determination of numbers of shared orthologues revealed highest numbers with *L. plantarum* (61–1 % of *L. salivarius* proteins), followed by *Enterococcus faecalis* (50–4 %), *L. sakei* (49·8 %), *L. acidophilus* (46·3 %) and *L. johnsonii* (45·1 %); other relevant genome features are shown in Table 1. The genomes of *L. plantarum* and *L. johnsonii* have been subjected to a detailed pairwise comparative analysis (Boekhorst et al., 2004). A total of 28 regions of conserved gene order were identified, totalling 0·75 Mb, but these regions were not collinear. Major differences were identified in reconstructed metabolic maps, biosynthetic capabilities and extracellular protein repertoires, providing a rational basis for directing experimental analyses. A focused comparative analysis of genome synteny and exopolysaccharide biosynthesis loci in *L. acidophilus, L. plantarum, L. johnsonii* and *Lactobacillus gasseri* has also been published (Klaenhammer et al., 2003), but exhaustive phylogenetic issues were not within the scope of that review.

The purpose of the present study was to determine the 16S rRNA gene phylogeny of the lactobacilli and to compare this with the data from several complementary whole-genome analytical approaches for the five publicly available *Lactobacillus* genomes. The phenotypic diversity of the genus was reflected in lack of genome backbone conservation, except between the closest species. Comparison of genome collinearity distances, functional categories of proteins, and analysis of a supergroup based on core proteins, has reinforced the genomic basis for the remarkable heterogeneity of this genus.

### METHODS

**Genome sequences.** The genomes analysed were the five complete and publicly available *Lactobacillus* genomes: *L. plantarum* WCFS1 (AL935263), *L. acidophilus* NCFM (CP000033), *L. johnsonii* NCC533 (AB017198), *L. sakei* subsp. sakei 23K (CR936503) and *L. salivarius* subsp. salivarius UCC118 (CP000233), including its 242 kb megaplasmid pMP118 (CP000234). The *L. salivarius* UCC118 megaplasmid was included because it constitutes 11 % of the total genome, and all *L. salivarius* strains harbour megaplasmids (Claesson et al., 2006), whereas plasmids are smaller and variably present in strains

<table>
<thead>
<tr>
<th></th>
<th><em>L. salivarius</em> UCC118 (pMP118)*</th>
<th><em>L. plantarum</em> WCFS1</th>
<th><em>L. acidophilus</em> NCFM</th>
<th><em>L. johnsonii</em> NCC533</th>
<th><em>L. sakei</em> 23K</th>
<th><em>E. faecalis</em> V583</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome size (bp)</td>
<td>1827 111 (242 436)</td>
<td>3308 274</td>
<td>1993 564</td>
<td>1992 676</td>
<td>1844 661</td>
<td>3218 031</td>
</tr>
<tr>
<td>G+C content (mol%)</td>
<td>32·9 (32·1)</td>
<td>44·4</td>
<td>34·7</td>
<td>34·6</td>
<td>41·3</td>
<td>37·5</td>
</tr>
<tr>
<td>Number of genes</td>
<td>1765 (242)</td>
<td>3051</td>
<td>1864</td>
<td>1821</td>
<td>1884</td>
<td>3113</td>
</tr>
<tr>
<td>Unique genes†</td>
<td>483 (146)/224 (80)</td>
<td>1289/412</td>
<td>503/181</td>
<td>453/90</td>
<td>487/230</td>
<td>1626/595</td>
</tr>
<tr>
<td>Gene density (%)</td>
<td>84·1 (75–6)</td>
<td>83·7</td>
<td>88·2</td>
<td>89·3</td>
<td>86·9</td>
<td>86·1</td>
</tr>
<tr>
<td>tRNAs</td>
<td>78 (0)</td>
<td>70</td>
<td>61</td>
<td>79</td>
<td>63</td>
<td>68</td>
</tr>
<tr>
<td>rRNA operons</td>
<td>7 (0)</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>IS elements</td>
<td>32 (11)</td>
<td>15</td>
<td>17</td>
<td>14</td>
<td>12</td>
<td>38</td>
</tr>
<tr>
<td>Pseudogenes</td>
<td>49 (20)</td>
<td>39</td>
<td>NA</td>
<td>NA</td>
<td>30</td>
<td>6</td>
</tr>
<tr>
<td>Phage-related genes</td>
<td>167 (1)</td>
<td>159</td>
<td>18</td>
<td>54</td>
<td>22</td>
<td>21</td>
</tr>
</tbody>
</table>

*Numbers in parentheses refer to pMP118.
†Unique genes among these six species/Unique genes that had no hits against the non-redundant protein database with BLASTP E-value <1×10⁻¹⁰.
of the other sequenced species. The genome of E. faecalis V583 (AE016830), an opportunist pathogen and natural inhabitant of the mammalian gastrointestinal tract, had previously been shown to contain a high number of genes that had orthologues in L. salivarius (Claesson et al., 2006), and was therefore included in genome comparisons as a non-Lactobacillus outgroup. Functional annotations including COG (clusters of orthologous group) categories were obtained from NCBI, with the exception of L. salivarius, for which COG annotations were retrieved from the ERGO bioinformatics suite.

**Predicted proteome comparisons.** Each of the six predicted proteomes was searched for orthologues against the other proteomes, where orthology between two proteins was defined as best bidirectional FASTA (Pearson, 2000) hits. Such hits were also required to have at least 30% identity over at least 80% of both protein sequences. While this is an arbitrary definition of orthology, it has been applied in many comparative analyses from the Sanger Institute (e.g. Parkhill et al., 2003). This orthology definition excluded hits between single-component proteins and fusion proteins. Proteins sharing the same putative orthologues with each and every one of the other genomes were identified, to generate a set of shared orthologues. This stringent approach was also applied to the five Lactobacillus genomes (i.e. E. faecalis excluded) to investigate genus-specific proteins. These were then further searched against the non-redundant database to check for other non-Lactobacillus hits. Those proteins showing no orthology with any of the other five genomes were regarded as unique.

**Extraction of shared core genes and functional annotation.** To reconstruct the phylogeny of lactobacilli, our analysis started with the identification of those core genes that had been most conserved over a long period through evolution, so that the core was defined as the set of genes shared in each studied genome. To determine the core genes and their corresponding protein sequences, we extracted the orthologues present in each of the respective predicted proteomes in a two-step procedure. First we compared each protein against all the other proteins (all-against-all) using BLAST (Altschul et al., 1997) (cutoff: E-value $1 \times 10^{-10}$) and clustered them in 'families' according to their homology using the BLAST results. This clustering of families was performed using MCL, a graph-theory-based Markov clustering algorithm (Van Dongen, 2000; http://micans.org/mcl/). Following this, we selected the families that contained a single protein member in each genome, thereby discarding paralogues that could bias our analyses. Protein sequences corresponding to genes coming from mobile elements such as phages and IS elements were discarded manually. An in-house Perl script was used to compare the resulting list of proteins with that obtained using protein-based FASTA pairwise genome comparisons; only those proteins found by both methodologies were selected for analysis. A final total of 354 core proteins was obtained following this methodology.

Functional COG category assignments were collated for the core proteins thus identified, in order to gain further biological understanding from differences and similarities observed between the genomes. In rare cases where individual protein COG category assignments differed for members of core protein sets (presumably due to borderline-confidence COG assignments), the category containing the majority of proteins was adopted.

**Construction of a 16S rRNA-based phylogenetic tree.** A complete phylogenetic analysis was performed using the 111 16S rRNA gene sequences corresponding to the Lactobacillus species available in public databases (this exceeds the number of validated species discussed above). A non-lactobacillus DNA sequence was also included in the 16S tree, derived from the Pedococcus pentosaceus type strain. Most of the sequences were extracted from the Ribosomal Database Project-II (Cole et al., 2005); type strain sequences were chosen from each species where available. Those sequences that were not already curated in this database were extracted from the EMBL database using the public Sequence Retrieval System (SRS) server at EBI (Zdobnov et al., 2002). The 16S rRNA sequence of the Bacillus subtilis type strain was added as outgroup, since previous analyses have suggested that enterococci and streptococci are too close to the lactobacilli to use as outgroups for 16S phylogeny (Heilig et al., 2002).

Sequences were aligned with CLUSTALW (Thompson et al., 1994) using default parameters, prior to manual removal of gap regions, which was then followed by building the phylogenetic tree using the method of maximum-likelihood implemented in PhyML (Guindon & Gascuel, 2003) with the general time-reversible (GTR) model.

**Whole-genome alignments between species at DNA and protein level.** Lactobacillus full genome sequence alignments were performed in an all-against-all comparison using MUMmer (Kurtz et al., 2004). Comparisons were performed at both DNA and protein level. For the L. salivarius genome, the megaplasmid sequence was concatenated in succession to its chromosome sequence. To complement our graphical analysis, a quantitative approach was designed based on the numerical data obtained using MUMmer. Distances were calculated based on a modification of the method applied by Chan et al. (2006) where the values of total length of regions aligned at protein level with an identity higher than 60% using the PROmer (part of the MUMmer package) algorithm results replaced the total mums lengths. The calculation of the distance between two genomes is the logarithm of the PROmer alignments normalized by the size of the shortest genome being compared as follows:

$$D_{AB} = -\log_2 \left( \frac{\text{Total promer alignments}_{AB}}{\operatorname{min}(\text{Length Genome Sequences}_{AB})} \right)$$

**Orthologue tree.** Another quantitative approach, based upon distances derived from orthologue repertoires, was undertaken to compare with the phylogeny described above. After removing paralogues and mobile elements from the orthologue sets, two phylogenetic trees were constructed; the first was based on orthologue counts, where distances $D_{AB}$ were calculated as the logarithm of shared orthologues between two genomes, normalized by the weighted mean of the total number of genes (Korbel et al., 2002):

$$D_{AB} = -\ln \left( \frac{\sqrt{(\text{all genes})_A^2 + (\text{all genes})_B^2}}{\sqrt{2 \times (\text{all genes})_A \times (\text{all genes})_B}} \times \left( \frac{\text{shared orthologues}}{\text{all genes}} \right) \right) _{AB}$$

(1)

The second tree was constructed by including the percentage identity and hit-length for each orthologue pair, rendering numbers of identical amino acids. The distances were then calculated as the logarithm of the sum of these identical amino acids, normalized by the weighted mean genome size where $N$ is the number of orthologues between A and B:

$$D_{AB} = -\ln \left( \frac{\sqrt{(\text{genome size})_A^2 + (\text{genome size})_B^2}}{\sqrt{2 \times (\text{genome size})_A \times (\text{genome size})_B}} \times \sum_{i}^{N} \left( \frac{\text{identical amino acids}}{\text{genome size}} \right)_{i} \right) _{AB}$$

(2)

Both neighbour-joining and UPGMA techniques were used to build trees. To correct for the possibly faulty assumption that evolutionary rate is constant through these lineages, the transformed distance method was applied to the distance values, with E. faecalis as outgroup, prior to the UPGMA construction.

http://mic.sgmjournals.org
Supertree construction. Using CLUSTALW, 354 protein families were aligned and phylogenetic trees were built using MULTIPHYL (Keane et al., 2006, http://www.cs.nuim.ie/distributed/multi phyll.php). The best supertree was found following a heuristic search of tree space using the Most Similar Supertree (dift) (Creevey et al., 2004), Maximum Quartet fit (qfit) and Maximum Split fit (sfit) analysis methods as implemented in Clann 2.0.2 (Creevey & McInerney, 2005). The final representation of the supertree was built by SplitsTree (Huson, 1998).

RESULTS AND DISCUSSION

General genome characteristics

Three of the five Lactobacillus genome sequences are relatively similar in size, just smaller than 2 Mb, while that of L. plantarum is closer to the genome size of E. faecalis (Table 1). L. salivarius has the smallest chromosome, but the overall genome size is 2.06 Mb because of the presence of the megaplasmid pMP118. L. plantarum WCF51 harbours three plasmids (1-9, 2-3 and 36 kb; van Kranenburg et al., 2005), and L. salivarius UCC118 harbours two plasmids of 20 kb and 44 kb (Flynn, 2001), while the other sequenced genomes lack plasmids, although extrachromosomal replicons are very common in lactobacilli (Wang & Lee, 1997). L. plantarum and L. sakei have a distinctively high GC content (above 40 mol%), whereas that of the other Lactobacillus genomes ranges from 32.9 to 34.7% (Table 1). There are no unusual species anomalies with respect to coding density, tRNA gene number or rRNA locus number, the latter being more consistent within the genus than, for example, the range of one to four rRNA operons found in another commensal, Bifidobacterium spp. (Candela et al., 2004). As previously noted (Claesson et al., 2006), the L. salivarius UCC118 genome has the highest number of pseudogenes and IS elements, elevated by the substantial contribution of the megaplasmid to both of these features. It also has the highest number of bacteriophage-related genes, although there are only two complete prophages and two large remnants (Ventura et al., 2006). L. plantarum has proportionally the lowest number of pseudogenes and IS elements (relative to genome size) among the sequenced lactobacilli. Consideration of individual IS elements in the lactobacilli revealed some present in the genomes of several different species, but the target sequences are not conserved (data not shown). No single pseudogene occurred in more than one genome, indicating that inactivation and deterioration of coding sequences seem to be entirely species-specific (or stochastic) among the genomes analysed.

Phylogeny based on the 16S rRNA gene and diversity of lactobacilli

Since the original description of the genus Lactobacillus by Beijerinck (Beijerinck, 1901; Skerman et al., 1980), more than 100 different species have been described. It is widely recognized that the taxonomy of this genus is unsatisfactory due to the highly heterogeneous nature of its members (Schleifer & Ludwig, 1995). Despite the significant increase in the number of recognized Lactobacillus species, a complete phylogeny has not been published since Dellaglio’s revision (Dellaglio et al., 2005). For this comparative genomic analysis, we elected to use 16S rRNA gene phylogenies as our baseline, as it is the most established marker. A maximum-likelihood phylogenetic tree was therefore built using 111 16S rRNA gene sequences from currently available Lactobacillus species entries. The resulting tree is depicted in Fig. 1. This analysis using the most classic molecular marker clearly depicts the biological diversity in lactobacilli; however, some order can also be observed. Based on the phylogeny reconstruction, it is possible to group lactobacilli in five main divisions by imposing the criteria of number of species (>10 species) or high bootstrap values. Some of the latter values are low at major divisions, as is typical when large numbers of taxa are analysed (Takahashi et al., 1999). Group A is the largest division, containing 28 species, and includes the established members of the L. acidophilus complex (Kullen et al., 2000). Species not present in any group thus correspond to branches with small numbers of species and with low bootstrap values.

Although there is no clear biological correlation for this grouping, a distinct ecological niche association can be observed for most members of groups: for groups A and E, there is a GI tract association, and for groups C and D, there is an association with the environment and food. Group B and the other small independent branches show no clear niche predominance. This proposed division is in accordance with the one of Hammes & Hertel (2003): the Lactobacillus delbrueckii group corresponds to Division A, the Lactobacillus reuterii group corresponds to Division B, the Lactobacillus buchneri group corresponds to Division C, the L. plantarum group corresponds to Division D and the L. salivarius group corresponds to Division E. However, the L. sakei group has too few members to be considered a full division according to our criteria and the Lactobacillus casei group, as pointed out by Dellaglio & Felis (2005), appears now as two separate small clusters.

Based on their taxonomy, Dellaglio has recently suggested reclassification of the lactobacilli (Dellaglio et al., 2005), proposing a division into eight groups. As noted, the presence of clear divisions can also be observed in our tree topology of the 16S-based phylogeny. However, a single phylogenetic marker may be insufficient for tree resolution due to bias or lack of fine resolution, favouring usage of several markers, as proposed by the International Committee on Systematics of Prokaryotes (ICSP) (Vandamme et al., 1996). We therefore applied a range of whole-genome-based approaches. Four of the species for which genome sequences were available fell into three of the five major 16S rRNA phylogeny groups: L. acidophilus and L. johnsonii in group A, L. plantarum in group D and L. salivarius in group E. Complete genome sequences for representatives of groups B and C were not available.
Whole-genome alignments

Our first approach was the alignment of whole genome sequences. Though some degree of relationships could be visualized at DNA level (data not shown), phylogenetic signals quickly disappear as DNA sequences diverge, as a reflection of the increasing distance between the lactobacilli. The fact that long-range genome alignments of *Lactobacillus* spp. could not be produced at a DNA level is a significant indication of profound diversity, contrasting for example with other DNA–DNA interspecific alignments such as those of *Mycobacterium* spp. (Brussow et al., 2004). Protein rather than DNA alignments were therefore employed for these interspecific genome comparisons, providing a better picture across the genus.

The degree of alignment between species (Fig. 2) varied depending on the phylogenetic distance of the genomes being compared. The sequences with the best alignment are *L. johnsonii* and *L. acidophilus*, the almost continuous line closest DNA–DNA alignments that could be achieved for the lactobacilli were between members of the proposed *Lactobacillus* groups, such as in group A between *L. acidophilus* and *L. johnsonii* (data not shown). Protein rather than DNA alignments were therefore employed for these interspecific genome comparisons, providing a better picture across the genus.

The degree of alignment between species (Fig. 2) varied depending on the phylogenetic distance of the genomes being compared. The sequences with the best alignment are *L. johnsonii* and *L. acidophilus*, the almost continuous line closest DNA–DNA alignments that could be achieved for the lactobacilli were between members of the proposed *Lactobacillus* groups, such as in group A between *L. acidophilus* and *L. johnsonii* (data not shown). Protein rather than DNA alignments were therefore employed for these interspecific genome comparisons, providing a better picture across the genus.

The degree of alignment between species (Fig. 2) varied depending on the phylogenetic distance of the genomes being compared. The sequences with the best alignment are *L. johnsonii* and *L. acidophilus*, the almost continuous line closest DNA–DNA alignments that could be achieved for the lactobacilli were between members of the proposed *Lactobacillus* groups, such as in group A between *L. acidophilus* and *L. johnsonii* (data not shown). Protein rather than DNA alignments were therefore employed for these interspecific genome comparisons, providing a better picture across the genus.

The degree of alignment between species (Fig. 2) varied depending on the phylogenetic distance of the genomes being compared. The sequences with the best alignment are *L. johnsonii* and *L. acidophilus*, the almost continuous line closest DNA–DNA alignments that could be achieved for the lactobacilli were between members of the proposed *Lactobacillus* groups, such as in group A between *L. acidophilus* and *L. johnsonii* (data not shown). Protein rather than DNA alignments were therefore employed for these interspecific genome comparisons, providing a better picture across the genus.

The degree of alignment between species (Fig. 2) varied depending on the phylogenetic distance of the genomes being compared. The sequences with the best alignment are *L. johnsonii* and *L. acidophilus*, the almost continuous line closest DNA–DNA alignments that could be achieved for the lactobacilli were between members of the proposed *Lactobacillus* groups, such as in group A between *L. acidophilus* and *L. johnsonii* (data not shown). Protein rather than DNA alignments were therefore employed for these interspecific genome comparisons, providing a better picture across the genus.

The degree of alignment between species (Fig. 2) varied depending on the phylogenetic distance of the genomes being compared. The sequences with the best alignment are *L. johnsonii* and *L. acidophilus*, the almost continuous line closest DNA–DNA alignments that could be achieved for the lactobacilli were between members of the proposed *Lactobacillus* groups, such as in group A between *L. acidophilus* and *L. johnsonii* (data not shown). Protein rather than DNA alignments were therefore employed for these interspecific genome comparisons, providing a better picture across the genus.

The degree of alignment between species (Fig. 2) varied depending on the phylogenetic distance of the genomes being compared. The sequences with the best alignment are *L. johnsonii* and *L. acidophilus*, the almost continuous line closest DNA–DNA alignments that could be achieved for the lactobacilli were between members of the proposed *Lactobacillus* groups, such as in group A between *L. acidophilus* and *L. johnsonii* (data not shown). Protein rather than DNA alignments were therefore employed for these interspecific genome comparisons, providing a better picture across the genus.

The degree of alignment between species (Fig. 2) varied depending on the phylogenetic distance of the genomes being compared. The sequences with the best alignment are *L. johnsonii* and *L. acidophilus*, the almost continuous line closest DNA–DNA alignments that could be achieved for the lactobacilli were between members of the proposed *Lactobacillus* groups, such as in group A between *L. acidophilus* and *L. johnsonii* (data not shown). Protein rather than DNA alignments were therefore employed for these interspecific genome comparisons, providing a better picture across the genus.

The degree of alignment between species (Fig. 2) varied depending on the phylogenetic distance of the genomes being compared. The sequences with the best alignment are *L. johnsonii* and *L. acidophilus*, the almost continuous line closest DNA–DNA alignments that could be achieved for the lactobacilli were between members of the proposed *Lactobacillus* groups, such as in group A between *L. acidophilus* and *L. johnsonii* (data not shown). Protein rather than DNA alignments were therefore employed for these interspecific genome comparisons, providing a better picture across the genus.

The degree of alignment between species (Fig. 2) varied depending on the phylogenetic distance of the genomes being compared. The sequences with the best alignment are *L. johnsonii* and *L. acidophilus*, the almost continuous line closest DNA–DNA alignments that could be achieved for the lactobacilli were between members of the proposed *Lactobacillus* groups, such as in group A between *L. acidophilus* and *L. johnsonii* (data not shown). Protein rather than DNA alignments were therefore employed for these interspecific genome comparisons, providing a better picture across the genus.

The degree of alignment between species (Fig. 2) varied depending on the phylogenetic distance of the genomes being compared. The sequences with the best alignment are *L. johnsonii* and *L. acidophilus*, the almost continuous line closest DNA–DNA alignments that could be achieved for the lactobacilli were between members of the proposed *Lactobacillus* groups, such as in group A between *L. acidophilus* and *L. johnsonii* (data not shown). Protein rather than DNA alignments were therefore employed for these interspecific genome comparisons, providing a better picture across the genus.

The degree of alignment between species (Fig. 2) varied depending on the phylogenetic distance of the genomes being compared. The sequences with the best alignment are *L. johnsonii* and *L. acidophilus*, the almost continuous line closest DNA–DNA alignments that could be achieved for the lactobacilli were between members of the proposed *Lactobacillus* groups, such as in group A between *L. acidophilus* and *L. johnsonii* (data not shown). Protein rather than DNA alignments were therefore employed for these interspecific genome comparisons, providing a better picture across the genus.

The degree of alignment between species (Fig. 2) varied depending on the phylogenetic distance of the genomes being compared. The sequences with the best alignment are *L. johnsonii* and *L. acidophilus*, the almost continuous line closest DNA–DNA alignments that could be achieved for the lactobacilli were between members of the proposed *Lactobacillus* groups, such as in group A between *L. acidophilus* and *L. johnsonii* (data not shown). Protein rather than DNA alignments were therefore employed for these interspecific genome comparisons, providing a better picture across the genus.
observed in the xy graph being an expression of the high level of synteny between these two genomes. Their high degree of similarity reflected by the PROmer alignment can also be seen in their proximity in the 16S-based tree, both belonging to group A in Fig. 1. In contrast, the least related genomes, based on their low degree of alignment, were the L. salivarius genome (group E), versus those of L. acidophilus and L. johnsonii (group A), which correlates well with the distance between these genomes observed in the 16S-based tree.

A striking observation is that most of the alignments obtained show much lower degrees of synteny at interspecies level than in other species genome comparisons in high- and low-G+C-content Gram-positive bacteria (Brussow et al., 2004). The alignments are made of patched regions following an X-shaped distribution that is symmetric about the origin of replication as previously noted in other bacteria by Eisen et al. (2000). This symmetry indicates that matching sequences tend to occur at the same distance from the origin but not necessarily on the same side of the origin, which is explained by the fork replication theory (Tillier & Collins, 2000). These X-shaped regions are not clearly symmetrical with respect to the origin–terminus axis in any of the L. plantarum alignments. It is noteworthy that these X-shaped alignments can be distinguished most clearly in the middle region of the xy plots. The leftmost and

---

**Fig. 2.** Between-species whole-genome alignments of Lactobacilli generated by PROmer. Lower-right corner of each plot corresponds to the origin of replication for each genome. Plotted points of red and green colour indicate the forward and reverse matching substring regions, respectively. Regions in L. salivarius above the dotted lines correspond to the megaplasmid; in L. plantarum regions to the right of the dotted lines correspond to regions that show no synteny with other lactobacilli.
rightmost parts of the plots that correspond to the ori region have no clear regions of synteny, which may indicate that these regions are subject to a greater rate of change by a higher rate of reshuffling or by an increased acquisition of foreign DNA by horizontal gene transfer. This contrasts with previous analyses of other bacteria (Daubin & Perriere, 2003), which suggested that the region opposite to the replication origin (terminus) had a higher frequency of phage insertion and ‘orphan’ gene acquisition. Other important regions with a null synteny correspond to the megaplasmid pMP118 and a large region next to the origin of replication at the end of the L. plantarum circular chromosome as depicted in Fig. 2. The overall conclusion of the PROmer alignments, however, is that there is a very significant lack of collinearity, emphasizing the genome diversity of the sequenced lactobacilli.

Quantitative data can also be extracted from genome alignments as demonstrated by Chan et al. (2006), who used the total lengths of the mums regions. Though the use of genome alignments at DNA level using MUMmer has been validated for reconstruction of broad phylogenetic relationships of prokaryotic genomes (Chan et al., 2006; Henz et al., 2005), it might lack sensitivity for reconstructing phylogenies if we limit our analysis at intrageneric level. Due to the relatedness of our genomes, we slightly modified the distance measure formula: the regions aligned using PROmer were used instead of using the mums lengths. Finally, a neighbour-joining tree was built using the calculated distances (Fig. 3). The position of E. faecalis relative to Streptococcus thermophilus and B. subtilis used to outgroup this tree highlights the close relationship of E. faecalis with the Lactobacillus group compared to other close genera, which in turn indicates the complexity of the number of genera that might be included within this group. The close relationship between L. johnsonii and L. acidophilus is still conserved in the PROmer-based tree. However, some discrepancies in the positions of other species with respect to the 16S tree are observed. L. salivarius is no longer one of the more external species in the tree topology, this position being now occupied by L. sakei. In this new tree, L. salivarius and L. plantarum share a common ancestor based on the regions aligned at protein level. A deeper analysis based on gene content and the corresponding predicted proteome in lactobacilli was therefore performed to complement and clarify these relationships.

**Lactobacillus orthologues and taxon-specific proteins**

Orthologues are genes/proteins that have been separated from a last common ancestor, in an evolutionary time-frame, by a speciation event (Fitch, 1970). We identified 593 orthologues shared by the five *Lactobacillus* species, a number that was reduced to 518 when *E. faecalis* was included in the analysis (Fig. 4a). The most numerous core proteins were those involved in housekeeping functions including information processing, replication, cell envelope biogenesis, carbohydrate metabolism and signal transduction. The category containing genes encoding proteins with unknown function was second largest, emphasizing the need for functional studies to determine the cellular role of these gene products in *Lactobacillus* biology.

Among the 75 *Lactobacillus* core proteins not shared with *E. faecalis*, 11 were involved in translation, 7 were involved in cell division, 12 were related to carbon and amino acid metabolism, while 21 had no predicted function. When these 75 proteins were used to search against the non-redundant protein database, which also contains draft sequences of *L. gasseri*, *L. delbrueckii* and *L. casei*, many non-*Lactobacillus* matches (using BLASTP *E*-value of $1 \times 10^{-10}$ as cutoff) were found. This indicates that they are not *Lactobacillus*-specific, but simply core proteins that are absent in *E. faecalis*. This somewhat unexpected lack of genus-specific proteins again highlights the substantial diversification of the lactobacilli.

A logical extension of the analysis of common characteristics was to investigate species-specific proteins. Between 30% and 50% of each of the six genomes (i.e. including *E.

---

**Fig. 3.** (a) Neighbour-joining tree of sequenced lactobacilli and *E. faecalis*, *S. thermophilus* and *B. subtilis* based on PROmer distances. The tree is based on the distances calculated from the non-overlapped protein aligned regions. (b) Orthology-based phylogeny of the sequenced lactobacilli and *E. faecalis*, *S. thermophilus* and *B. subtilis*. The tree was constructed from neighbour-joining distances, based on the number of shared orthologues, normalized by the weighted mean of the total number of genes.
faecalis) analysed showed no significant similarity to the other genomes, and the corresponding proteins were thus considered unique. Fig. 4(b) shows the COG categories of unique proteins as a percentage of the total number of proteins for each species. The largest category represents proteins of unknown function. It is also evident that the two largest genomes (L. plantarum and E. faecalis) have a much higher percentage of unique proteins involved in transcription and signal transduction. The correlation of increased genome size and increased number of proteins in these categories has been previously noted for Pseudomonas (Stover et al., 2000). This probably reflects a requirement for more complex regulation of genome expression, as metabolic and cellular complexity increases proportional to genome size.

Certain other functional categories in bacterial genomes have previously been described to correlate either positively or negatively with genome size (Konstantinidis & Tiedje, 2004; Ranea et al., 2004). Our analysis of the six genomes identified such categories that display the same pattern for the two largest genomes, namely transport and metabolism of carbohydrates, amino acids and secondary metabolites.

Gene duplication is postulated to have played a major role in the evolution of biological novelty and biodiversity (Hancock, 2005). Such events within the respective genomes were elucidated by investigating and comparing paralogue numbers. Some examples where one genome has a significantly higher number of genes with the same or similar function are DNA-binding response regulators (11 in E. faecalis, 5–6 in others); β-glucoside phosphotransferase system (PTS) enzyme IIABC (13 in L. johnsonii, 2–8 in others); PTS system, IID component (13 in E. faecalis, 1–5 in others); multidrug transport proteins (10 in L. johnsonii, 2–5 in others); ABC transporter, ATP-binding proteins (10 in L. johnsonii, 2–5 in others); and maltose O-acetyltransferase (7 in L. johnsonii, 0–3 in others). Thus, gene duplications appear to be most frequent in E. faecalis and L. johnsonii, which is surprising for the latter genome, since it is relatively small. Targeted analysis of these paralogues may explain the selective pressure for their amplification in the respective species and test if they are niche-specific adaptations.

Since E. faecalis is an opportunistic pathogen that may be vancomycin resistant, we also investigated if related antibiotic-resistance genes were present in the genomes of the commensals. The vancomycin-resistance region (EF1955–EF1963) of E. faecalis contains genes for an enolase, a triosephosphate isomerase and a phosphoglycerate kinase (EF1961–EF1963), which are present as a conserved arrangement in all the Lactobacillus genomes except L. acidophilus. Two additional sets of adjacent genes in the pathogenicity island of E. faecalis, ornithine cyclodeaminase with two hypothetical proteins (EF0616–EF0618), and DNA-damage-inducible protein J with one hypothetical protein (EF0512–EF0513), are present in the genome of L. salivarius (LSL_1273–LSL_1275 and LSL_0030–LSL_0030b, respectively). Based upon the current annotations, therefore, only metabolism genes or

Fig. 4. Common and specific functions based on COG categories. (a) Number of core genes with and without E. faecalis in the analysis; (b) percentage of unique genes for each species. A line in the bar for L. salivarius also indicates the fraction of genes located on pMP118. *Omitted to increase clarity of remaining groups; L. acidophilus and L. johnsonii (~15 %), L. plantarum, L. salivarius, L. sakei and E. faecalis (~20–25 %).
stress-resistance genes from this *E. faecalis* vancomycin-resistance locus are present in the lactobacilli.

The genome sizes of the sequenced lactobacilli are fairly homogeneous, with the exception of *L. plantarum*, whose genome is 60–75% larger than the others. A comparison of regions with anomalous G+C content and interruptions of synteny revealed only a few likely sites where insertions in the *L. plantarum* genome relative to the other four lactobacilli may have occurred (data not shown). These regions make up a total of approximately 380 kb, which fails to explain the considerable difference in size. Furthermore, no extensive gene duplications were found in the *L. plantarum* genome. Obvious deletions in the other *Lactobacillus* genomes are harder to distinguish due to the low degree of gene conservation among them. A study of the top BLAST hits of *L. plantarum* proteins against the non-redundant database showed that hits against the draft *P. pentosaceus* ATCC 25745 genome (http://www.jgi.doe.gov/) were most abundant, with 434 top matches, followed by *L. salivarius* (329), *L. sakei* (280) and *L. casei* ATCC 334 (229).

It is theoretically possible that *P. pentosaceus* is a major contributor of laterally acquired genes in *L. plantarum*, but lineage-specific gene loss in the other lactobacilli is a more plausible explanation for genome size differences. The imminent publication of *P. pentosaceus* and other lactobacilli genomes will shed light on genome plasticity and diversity among these related species.

Interestingly and on a different note, there were four genes present in the core that are harboured by the megaplasmid pMP118 of *L. salivarius* alcohol dehydrogenase (LSL_1901), L-serine dehydratase alpha subunit (LSL_1931), ribulose-5-phosphate-3-epimerase (LSL_1948) and ribose 5-phosphate isomerase (*L. salivarius* LSL_1806). The presence of these core genes in pMP118 may be evidence of a process of mobilization between the megaplasmid and the chromosome of vertically evolved genes, since no other core gene were found in the megaplasmid.

**Lactobacillus phylogeny and phylogenomics**

To complement the sequence-based and synteny-based phylogeny studies, we also constructed a distance-based phylogeny on the basis of orthologous gene content. Although not calculated from sequence-derived distances, trees of orthologue-number distances are based on the principle that the number of genes two genomes have in common depends on their evolutionary distance. These distances can be seen as reflecting evolutionary events such as gain and loss of genes, whereas the underlying properties such as the gene content can be interpreted in terms of function (Snel et al., 1999). Equations 1 and 2 (see Methods) were applied to the numbers of orthologue pairs and fractions of identical protein sequences for these pairs. Table 2 shows the number of orthologues for the 15 possible comparisons between the six species, as well as the mean percentage identity for these pairs, as another measure of similarity. Neighbour-joining trees were constructed based on these values. When the neighbour-joining tree was rooted with *E. faecalis*, *S. thermophilus* and *B. subtilis* as outgroups (Fig. 3b), it showed the same topology and similar evolutionary distances for both equations. Reassuringly, it was also in concordance with the topology displayed by a transformed UPGMA tree constructed from the same distance matrix (data not shown). The phylogeny based on orthology-derived distance values is also strikingly similar to that based upon PROmer-derived distances (Fig. 3a), confirming *L. sakei* as being the most peripheral of the sequenced species. In this respect it contradicts the 16S rRNA gene phylogeny. However, some discrepancies are found for the positions of *S. thermophilus* and *B. subtilis* in the PROmer-based tree, indicating perhaps the limitation of the PROmer approach to reconstruct the phylogeny if it goes beyond the genus level.

**Supertree analysis**

Gene trees have been important in phylogeny reconstruction. With the total gene content determination in the genomic era, a multigene tree phylogeny reconstruction approach provides a solution to the problem of combining evidence from different genome loci to infer phylogeny without losing information from independent gene histories. Three hundred and fifty-four maximum-likelihood phylogenetic trees corresponding to each of the core proteins previously identified were used to build a

---

**Table 2. Orthologue comparisons between Lactobacillus spp. and E. faecalis**

Paralogous genes and mobile elements were excluded from the analysis. The upper-right half of the matrix shows mean percentage residue identities between the protein sequences encoded by orthologous genes in the respective species, and the lower-left half shows numbers of orthologous genes (defined as described in Methods).

<table>
<thead>
<tr>
<th></th>
<th><em>L. salivarius</em></th>
<th><em>L. plantarum</em></th>
<th><em>L. acidophilus</em></th>
<th><em>L. johnsonii</em></th>
<th><em>L. sakei</em></th>
<th><em>E. faecalis</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>L. salivarius</em></td>
<td>X</td>
<td>58.5</td>
<td>54.1</td>
<td>54.3</td>
<td>57.8</td>
<td>56.1</td>
</tr>
<tr>
<td><em>L. plantarum</em></td>
<td>950</td>
<td>X</td>
<td>52.5</td>
<td>52.7</td>
<td>57.4</td>
<td>53.9</td>
</tr>
<tr>
<td><em>L. acidophilus</em></td>
<td>723</td>
<td>689</td>
<td>X</td>
<td>68.9</td>
<td>53.8</td>
<td>51.7</td>
</tr>
<tr>
<td><em>L. johnsonii</em></td>
<td>707</td>
<td>687</td>
<td>911</td>
<td>X</td>
<td>54.2</td>
<td>52.0</td>
</tr>
<tr>
<td><em>L. sakei</em></td>
<td>794</td>
<td>876</td>
<td>659</td>
<td>667</td>
<td>X</td>
<td>56.2</td>
</tr>
<tr>
<td><em>E. faecalis</em></td>
<td>816</td>
<td>879</td>
<td>631</td>
<td>660</td>
<td>863</td>
<td>X</td>
</tr>
</tbody>
</table>

---

http://mic.sgmjournals.org 3193

Downloaded from www.microbiologyresearch.org by
IP: 54.70.40.11
Lactobacillus supertree. Three methods available in Clann were used: the Most Similar Supertree Method (dfit), the Maximum Quartet fit (qfit) and the Maximum Split fit (sfit). Using the 354 trees, heuristic searches of supertree space were carried out and a single optimal supertree topology was identified using the three methods. In the supertree, L. sakei is closer to L. johnsonii and L. acidophilus while E. faecalis is closer to L. salivarius and L. plantarum than the other lactobacilli included in our analysis. Heuristic searches are blind hill-walking algorithms. They try to find the optimum answer to a problem without trying all possible answers. One hundred bootstrap resamplings of the input trees were carried out. High bootstrap value (>90) were only present for the node corresponding to L. acidophilus and L. johnsonii (data not shown). The same data were analysed using the consensus tree reconstruction in SplitsTree resulting in the tree shown in Fig. 5, which has the same topology as a tree obtained with Clann (not shown). Given the diversity and the high number of species in this bacterial genus, the closeness of L. salivarius and L. plantarum to E. faecalis may simply reflect their phylogenetic position at the edge of a very diverse group. Each of the four main branches in our tree (Fig. 5) may support the idea that these branches constitute taxa at the same level, with each branch a potential new genus.

To further explore this hypothesis, we re-examined the number of orthologues unique within the four divisions (Fig. 6). By this analysis L. acidophilus and L. johnsonii, well established as very related, shared 145 unique proteins, whereas L. salivarius and L. plantarum shared only 59 unique proteins. As expected from the tree topology, E. faecalis and L. sakei had relatively high numbers of unique proteins. Almost half of the COG assignments for the L. acidophilus–L. johnsonii unique proteins were ‘unknown function’, the majority of the remainder being carbon and amino acid metabolism, cell envelope biogenesis, and transcription (data not shown). The proportional distribution in the L. salivarius–L. plantarum division was similar. It seems logical that proteins uniquely present in the proposed species groups contribute to adaptations that typify these groups, and proteins identified in this analysis will be used for a targeted study to test this theory.

Concluding remarks
Analysis of whole genomes of the lactobacilli has provided novel evidence for species relatedness, and highlighted the possibility of creating subgeneric divisions based on total genome content. On-going and imminent genome sequencing projects of other lactobacilli will extend and consolidate this comparative framework. However, it is already apparent that some lactobacilli are more closely related to E. faecalis than they are to other lactobacilli, and other non-lactobacilli such as P. pentosaceus will likely cluster within or close to the genus when their genome sequence is completed and analysed. Using a set of core genes common to the five Lactobacillus genomes examined here, several phylogenetic techniques employed in this study produced clear divisions that are candidate novel taxa. We note that there is currently a relative overrepresentation of genome sequence information from small-genome nutritionally fastidious lactobacilli. Of the eight ongoing lactobacillus genome projects (http://www.genomesonline.org) the whole genome sequences of Lactobacillus brevis, L. casei, L. reuteri, Lactobacillus rhamnosus and also P. pediococcus will particularly help in elucidating the genomic diversity of the genus. A more strategic selection of future genome sequence targets will be useful for addressing remaining questions about gene loss and gene acquisition to explain variations in genome size, and for substantiating novel divisions of this diverse genus.

Fig. 5. Phylogenetic supertree based on the sequences of 354 Lactobacillus core proteins, using SplitsTree. Dotted lines show the proposed taxa.

Fig. 6. Venn diagram showing numbers of unique proteins for the four proposed divisions: L. salivarius–L. plantarum (Lsl-Lpl); L. acidophilus–L. johnsonii (Lac-Ljo); L. sakei (Lsa) and E. faecalis (Efa).
NOTE ADDED IN PROOF

The genome sequence of L. bulgaricus has recently been published (van de Guchte et al., Proc Natl Acad Sci U S A 103, 9274–9279). This species is part of the L. acidophilus complex and is therefore likely to cluster with L. acidophilus and L. johnsonii in the analyses described herein.

ACKNOWLEDGEMENTS

This research was supported by Science Foundation Ireland through a Centre for Science, Engineering and Technology award to the Alimentary Pharmabiotic Centre and by grants from the Higher Education Authority PRTLI1 and PRTLI3 programmes, the Department of Agriculture and Food FIRM 01/R&D/C/159 programme, and the Irish Research Council for Science, Engineering and Technology EMBARK postdoctoral programme (to C. C.). We thank J. Parkhill, R. Doolittle, J. McInerney and P. J. Lockhart for advice on various phylogenetic methodologies, and Yin Li and Marco Ventura for critical reading of the manuscript.

REFERENCES


Boekhorst, J., Siezen, R. J., Zwaalen, M. C. & 7 other authors (2004). The complete genomes of Lactobacillus plantarum and Lactobacillus johnsonii reveal extensive differences in chromosome organization and gene content. Microbiology 150, 3601–3611.


Dunne, C., O’Mahony, L., Murphy, L. & 11 other authors (2001). In vitro selection criteria for probiotic bacteria of human origin correlation with in vivo findings. Am J Clin Nutr 73, 3865–3925.


McKerthy, J., O'Mahony, L., O'Callaghan, L. & 8 others authors (2003). Double blind, placebo controlled trial of two probiotic strains in interleukin 10 knockout mice and mechanistic link with cytokine balance. *Gut* 52, 975–980.


