The complete genomes of *Lactobacillus plantarum* and *Lactobacillus johnsonii* reveal extensive differences in chromosome organization and gene content

Jos Boekhorst,[1] Roland J. Siezen,[1,2,4] Marie-Camille Zwahlen,[3] David Vilanova,[3] Raymond D. Pridmore,[3] Annick Mercenier,[3] Michiel Kleerebezem,[2,4] Willem M. de Vos,[2] Harald Brüssow[3] and Frank Desiere[3]

[1]Center for Molecular and Biomolecular Informatics, 6525ED, Nijmegen, The Netherlands

[2]Wageningen Centre for Food Sciences, 6700 AN Wageningen, The Netherlands

[3]Nestlé Research Center, Nestec SA, 1000 Lausanne 26, Switzerland

[4]NIZO food research, 6710 BA Ede, The Netherlands

Correspondence
Jos Boekhorst
J.Boekhorst@cmbi.kun.nl

Received 9 June 2004
Revised 23 July 2004
Accepted 29 July 2004

The first comprehensive comparative analysis of lactobacilli was done by comparing the genomes of *Lactobacillus plantarum* (3·3 Mb) and *Lactobacillus johnsonii* (2·0 Mb). *L. johnsonii* is predominantly found in the gastrointestinal tract, while *L. plantarum* is also found on plants and plant-derived material, and is used in a variety of industrial fermentations. The *L. plantarum* and *L. johnsonii* chromosomes have only 28 regions with conservation of gene order, totalling about 0·75 Mb; these regions are not co-linear, indicating major chromosomal rearrangements. Metabolic reconstruction indicates many differences between *L. johnsonii* and *L. plantarum*: numerous enzymes involved in sugar metabolism and in biosynthesis of amino acids, nucleotides, fatty acids and cofactors are lacking in *L. johnsonii*. Major differences were seen in the number and types of putative extracellular proteins, which are of interest because of their possible role in host–microbe interactions. The differences between *L. plantarum* and *L. johnsonii*, both in genome organization and gene content, are exceptionally large for two bacteria of the same genus, emphasizing the difficulty in taxonomic classification of lactobacilli.

INTRODUCTION

Lactobacilli belong to the lactic acid bacteria (LAB) and are members of the low-GC content Gram-positive bacteria. Many are used in starter cultures for food and feed fermentations, and several species are frequently encountered in the human gastrointestinal tract (Vaughan *et al.*, 2002). Some strains of LAB are marketed as probiotics, which are claimed to positively affect human and/or animal health (Braun-Fahrlander *et al.*, 2002; Link-Amster *et al.*, 1994). However, not much is known about the mechanisms by which these LAB affect the host.

Recently, the genomes of two members of the genus *Lactobacillus* have been completely sequenced: *Lactobacillus plantarum* WCFS1 (Kleerebezem *et al.*, 2003) and *Lactobacillus johnsonii* NCC533 (Pridmore *et al.*, 2004). *L. johnsonii* NCC533, isolated from human faeces, has been extensively studied for its probiotic activities, including immunomodulation (Haller *et al.*, 2000a, 2000b) and interaction with the human host (Ibnou-Zekri *et al.*, 2003). *L. plantarum* WCFS1 was isolated from human saliva. *L. plantarum* is a versatile bacterium that is found in a variety of ecological niches, ranging from vegetable and plant fermentations to the human gastrointestinal tract.

Abbreviations: BDI, base deviation index; CDS, coding sequence; COG, Clusters of Orthologous Group; KEGG, Kyoto Encyclopedia of Genes and Genomes; LAB, lactic acid bacteria; PTS, phosphotransferase system.

Details of the size and location of conserved gene clusters in *L. plantarum* and *L. johnsonii* may be found in Supplementary Table S1; the number of proteins of *L. plantarum* and *L. johnsonii* for all COG classes in Supplementary Table S2; a KEGG comparison of major differences between *L. plantarum* and *L. johnsonii* in Supplementary Table S3; *L. johnsonii* and *L. plantarum* API 50 test results in Supplementary Table S4; the redundancy of enzymes involved in pyruvate metabolism in Supplementary Table S5; gene clusters encoding functionally related proteins present in *L. plantarum* but not in *L. johnsonii* and vice versa in Supplementary Table S6; lists of proteins unique to either *L. plantarum* or *L. johnsonii* in Supplementary Tables S7–S12; lists of proteins involved in the biosynthesis of polysaccharides, bacteriocins and prophages in Supplementary Table S13 with the online version of this paper at http://mic.sgmjournals.org.

This flexibility of *L. plantarum* is reflected by its relatively large genome size, a large number of proteins involved in regulation and transport functions, and a high metabolic potential (Kleerebezem *et al.*, 2003).

In order to expand our understanding of the molecular evolution, diversity, function and adaptation of lactobacilli to specific environments, we have performed a whole-genome comparison of *L. plantarum* and *L. johnsonii*. In addition, we compared the proteins of these two organisms to the draft sequences of other LAB genomes (Klaenhammer *et al.*, 2002). We provide a first comprehensive view of differences on the genome level in lactobacilli, and evidence for large genetic diversity in this genus. We identify features underlying the large difference in genome size and gene content in lactobacilli, and provide a first insight into the set of genes and functions which could be specific for lactic acid bacteria. This knowledge provides numerous leads for targeted experimental verification of unique or common physiological properties.

METHODS

Genome sequences. Complete genome sequences of *L. plantarum* WCFS1 (Kleerebezem *et al.*, 2003) accession number AL935263, *L. johnsonii* NCC533 (Pridmore *et al.*, 2004) accession number AE017198, *Bacillus subtilis* 128, *Enterococcus faecalis* V583, *Listeria monocytogenes* EGDe and *Lactococcus lactis* IL1403 were obtained from GenBank Entrez Genomes (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html). The gene ID numbers used in the text to refer to specific *L. plantarum* and *L. johnsonii* genes are the same as those used in the original papers (Kleerebezem *et al.*, 2003; Pridmore *et al.*, 2004). Genome comparison with unfinished LAB genome sequences utilized sequence data of *Lactobacillus brevis* ATCC367, *Lactobacillus casei* ATCC334, *Lactobacillus delbrueckii* ATCCBAA-365, *Lactobacillus gasseri* ATCC-33323, *Leuconostoc mesenteroides* ATCC-8293, *Oenococcus oeni* PSU-1 and *Pediococcus pentosaceus* from the ERGO database (http://ergo.integratedgenomics.com/ERGO/), originally produced by the US Department of Energy Joint Genome Institute (http://www.jgi.doe.gov). For comparative purposes, all species of the genus *Lactobacillus*, *Lactococcus* and *Leuconostoc*, as well as the bacteria *Streptococcus thermophilus*, *Oenococcus oeni*, *Bifidobacterium longum* and *Pediococcus pentosaceus*, are considered to be LAB (Klaenhammer *et al.*, 2002). Fig. 1 shows the 16S rRNA tree of the relevant organisms.

The coding sequences (CDSs) of the *L. johnsonii* genome have been identified using FrameD (Schiex *et al.*, 2003), while the CDSs of the *L. plantarum* genome have been identified using Glimmer (Delcher *et al.*, 1999), which could lead to some erroneous comparison of the CDSs. However, for both organisms the positions of CDSs on the genome have been manually adjusted based on the presence of a plausible ribosome-binding site and on BLAST alignments with homologues (Kleerebezem *et al.*, 2003; Pridmore *et al.*, 2004), reducing the impact of this difference in CDS identification. Moreover, for both organisms the minimal size of a CDS was set at 30 codons.

Genome comparisons. Orthologous relationships were detected by a previously described method (Snel *et al.*, 2002) using the Smith & Waterman sequence comparison algorithm (Smith & Waterman, 1981) against the NCBI Clusters of Orthologous Group (COG) database (Tatusov *et al.*, 2001). The functional classification provided by the COG database was used for the functional comparison of *L. plantarum* and *L. johnsonii* on a genome-wide scale.

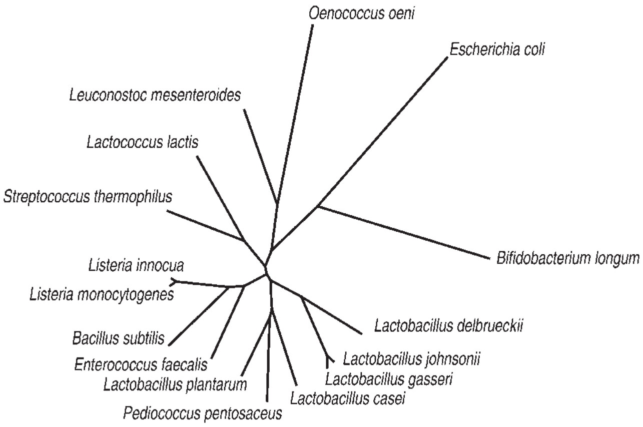

Fig. 1. 16S rRNA-based phylogenetic tree (unrooted). Sequences were extracted from the European rRNA database (Wuyts *et al.*, 2004) and aligned using CLUSTAL W (Thompson *et al.*, 1994). The tree was visualized using TreeView (Page, 1996).

Homology relationships were established using BLASTP (Altschul *et al.*, 1990) and Smith & Waterman sequence comparison. Homologues were detected with a threshold of $1E^{-10}$; a gene was considered organism specific when it had no Smith & Waterman hits at all, or only hits with an e-score higher than $1E^{-10}$ to proteins of other organisms in the non-redundant proteins databases (SWISS-PROT, TrREMBL and TrEMBL updates) (Boeckmann *et al.*, 2003) or the LAB genomes taken from the ERGO database. Proteins were considered LAB-specific when they did not have a Smith & Waterman hit with an e-score lower than $1E^{-10}$ in a search against SWISS-PROT, TrREMBL, TrEMBL updates and the LAB sequences taken from the ERGO database.

Whole genomes were compared at the nucleotide level using the Dotter software (Sonnhammer & Durbin, 1995) with default values. A bidirectional best-hit approach was used to identify genome synteny at the protein level. The results of this analysis were visualized using the Artemis Comparison Tool (http://www.sanger.ac.uk/Software/ACT/).

Transporter classification was preformed according to the TC-DB scheme (Busch & Saier, 2002). All proteins were searched against the TC-DB Database Release 1.5.1 using BLASTP with a threshold of $10E^{-4}$, followed by manual curation: false positive hits were removed manually when clear evidence suggested that they were not related to transport function.

Signal peptides were predicted using SignalP (Nielsen *et al.*, 1997).

Base deviation analysis of genes was performed by calculating a chi-squared index based on the expected and observed frequency for each nucleotide (Tettelin *et al.*, 2001).

Synchronizing annotation. The two genomes compared in this study were initially analysed using different ontologies and annotations (Kleerebezem *et al.*, 2003; Pridmore *et al.*, 2004). To facilitate functional comparison of *L. plantarum* and *L. johnsonii*, the annotation of proteins found to be homologous, but having different annotations in the two genomes, was manually verified and corrected where necessary. This resulted in an improved annotation of both genomes, in particular for the functional class 'regulation' and for the assignment of EC numbers, and made automated detection of functional differences possible.

Table 1. Genome features of *L. plantarum* WCFS1 and *L. johnsonii* NCC533

Genome feature	Strain	
	L. plantarum WCFS1	*L. johnsonii* NCC533
Length (bp)	3 308 274	1 992 676
Coding density (%)	84·1	89·3
G+C content (%)	45·6	34·9
Predicted CDSs	3 009	1 821
tRNAs	62	79
rRNA operons	5	6
Phage genes	159 (2 prophages, 2 remnants)	54 (2 prophages, 1 remnant)
IS elements	15	14

Reconstruction of metabolic pathways. EC numbers were extracted from the genome annotations and manually curated. They were then automatically mapped onto the Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathways (Kanehisa *et al.*, 2002) for visualization and identification of differences in metabolism between *L. plantarum* and *L. johnsonii.* In cases of predicted missing key enzymes in one of the two organisms, a further effort was made to identify homologous candidate enzymes by extensive manual searches with BLASTP and HMMER (Eddy, 1996; Sonnhammer *et al.*, 1998).

Sugar utilization. API 50 analysis of sugar utilization was performed using the supplier's protocol (BioMérieux Benelux). Additional sugar fermentation profiles were obtained from the literature (Fujisawa *et al.*, 1992; Kleerebezem *et al.*, 2003).

RESULTS AND DISCUSSION

General genome features

The main features of the genomes of *L. plantarum* WCFS1 and *L. johnsonii* NCC533 are shown in Table 1. *L. plantarum* WCFS1 has a genome of over 3·3 Mb, which is exceptionally large for a *Lactobacillus* species, since the genome size of lactobacilli is generally between 1·8 and 2·5 Mb (Klaenhammer *et al.*, 2002). At the DNA level, *L. johnsonii* and *L. plantarum* are very divergent. A DNA dot plot comparison of the two genomes (data not shown) shows a low overall sequence similarity. Much closer DNA homology has been observed between *L. johnsonii* and other members of the *Lactobacillus acidophilus* family (Pridmore *et al.*, 2004).

Homologous proteins

Evolutionary distances can be measured by the comparison of gene repertoires (Tamames, 2001). Closely related species share a large proportion of genes; in contrast, distantly related species should have lost a significant fraction of the genes inherited from their last common ancestor, resulting in a low proportion of shared genes. An overview of the percentage of homologues shared between the various genomes is given in Table 2. Of all the proteins encoded by the *L. johnsonii* genome, 83 % have a homologue in *L. gasseri*, 70 % have a homologue in *L. plantarum*, 62 % have a homologue in *E. faecalis* and 58 % have a homologue in *L. monocytogenes*. In contrast, when the *L. plantarum* genome is used as query, the large difference in genome size between *L. plantarum* and *L. johnsonii* leads to *L. plantarum* sharing more homologues with the larger genomes of *E. faecalis* (58 %), *L. monocytogenes* (57 %) and *B. subtilis* (52 %), than with *L. johnsonii* (51 %).

Table 2. Homologous proteins in genomes of Gram-positive bacteria

The numbers indicate the percentage of the proteins in the query genome with a homologue (BLAST hit with e-score higher than $1E^{-10}$) in the other genome. Amino acid sequences of the species in the header row are used as database, the sequences of the species in the column as query.

Species	Species						
	1	2	3	4	5	6	7
1 *B. subtilis*	–	43	29	30	40	37	49
2 *E. faecalis*	51	–	40	41	53	52	56
3 *L. gasseri*	53	63	–	85	69	59	58
4 *L. johnsonii*	54	63	83	–	70	60	58
5 *L. plantarum*	52	58	49	51	–	54	57
6 *L. lactis*	52	60	44	46	59	–	58
7 *L. monocytogenes*	65	60	42	44	58	53	–

Genome synteny

On an evolutionary time scale, protein sequences are more conserved than DNA sequences. It is possible to detect gene clusters encoding homologous proteins in related organisms even where low-level DNA conservation makes sequence alignment very difficult. These syntenic regions can provide insight into functions of the proteins comprising them: for example, genes already described in one organism might be annotated correctly in a second organism based on synteny. This principle has been used in the prediction of gene function by several methods, such as Rosetta Stone (Marcotte *et al.*, 1999) and the conserved gene neighbours method (Dandekar *et al.*, 1998; Overbeek *et al.*, 1999). The selective advantage of physical proximity of genes for co-regulation makes some gene clusters less prone to breakup than others, thus extending the range of evolutionary distance over which sequence conservation is detectable.

A dot plot comparison at the protein level of the genomes of *L. plantarum* and *L. johnsonii* (Fig. 2) shows no large-scale conservation of gene order, but only conservation of genes in clusters, confirming the relatively large phylogenetic distance between *L. plantarum* and *L. johnsonii.* The lack of large-scale gene order conservation between *L. plantarum* and *L. johnsonii* is in strong contrast to the whole chromosome alignment of *L. johnsonii* and *L. gasseri*,

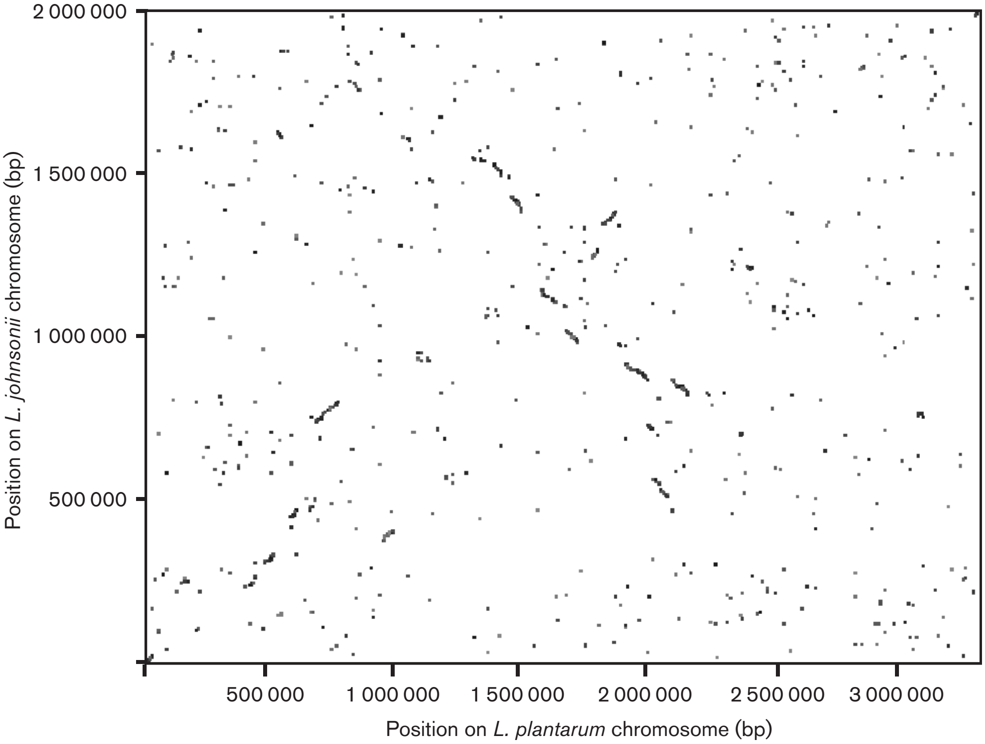

Fig. 2. Whole-genome protein comparison of *L. johnsonii* NCC533 and *L. plantarum* WCFS1 using BLASTP. Dots represent homologous proteins. The position of a dot represents the position of the homologous proteins on the *L. johnsonii* and *L. plantarum* chromosomes. Base numbering starts at the origin of replication. The hits were limited to those of >50 % protein alignment length.

which shows a high degree of conservation and synteny over the whole genome (Pridmore *et al.*, 2004). *L. johnsonii* and *L. plantarum* share only 28 large regions of conserved gene order, ranging in size from 7 (arbitrarily defined as minimum) to 75 genes, and encoding nearly 550 conserved proteins. Details of the size and location of these clusters may be found in Supplementary Table S1 with the online version of this paper at http://mic.sgmjournals.org.

Although the order of the orthologous genes in these clusters is conserved, some of the clusters contain insertions in one of the two bacterial chromosome sequences. Fig. 3 shows an example of such a cluster in which some of the genes unique for *L. plantarum* are found inserted in a conserved cluster. In ten of the conserved clusters, most genes in the cluster are functionally related (Supplementary Table S1), while the residual clusters contain genes that encode proteins involved in different cellular functions. The former clusters encompass the well-documented Nus-A/Inf-B cluster (Shazand *et al.*, 1993) and the macromolecular synthesis cluster (Metzger *et al.*, 1994). Most of the 28 clusters correspond to regions of protein sequence conservation across genus borders in Gram-positive bacteria, as many clusters are also found in the *B. subtilis*, *E. faecalis* and *L. monocytogenes* genomes (data not shown). This low degree of synteny between *L. plantarum* and *L. johnsonii* suggests that they are only marginally more related to each other than to the other Gram-positive bacteria.

Clusters of orthologous genes conserved between *L. plantarum* and *L. johnsonii* are located near the diagonals between the origin and terminus of replication showing a weak X-alignment pattern between the chromosomes (see Fig. 2). This observation indicates multiple chromosomal inversions pivoted on the terminus and origin of replication, causing major rearrangements. It has been suggested that this phenomenon is mostly caused by recombination occurring between, or close to, replication forks (Tillier & Collins, 2000). The degree of synteny can be related to the phylogenetic distance of the organisms: closer genomes have a more distinct X-alignment than more distant genomes (Suyama & Bork, 2001).

A similar synteny analysis for *L. johnsonii* NCC533 or *L. plantarum* WCFS1 with *E. faecalis* showed lower conservation, but many of the same clusters could be identified (data not shown). The number of conserved clusters, as well as the number of syntenic genes in the clusters, is smaller than in the *johnsonii*/*plantarum* comparison, but the degree of overall conservation corroborated well the fact that the genus *Enterococcus* is closely related to but distinct from *Lactobacillus* (Klein, 2003). Very limited

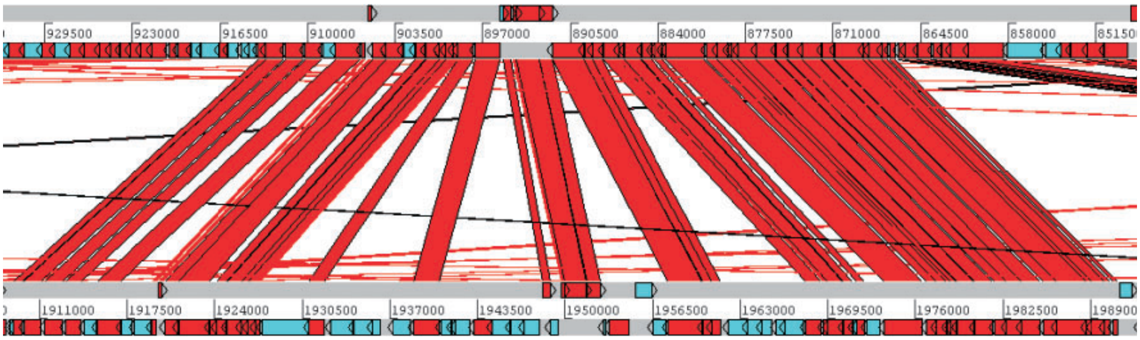

Fig. 3. Example of a gene cluster conserved between *L. johnsonii* and *L. plantarum*. The two horizontal bars at the top represent the two strands of the *L. johnsonii* genome; the two at the bottom represent the two strands of the *L. plantarum* genome. Horizontal arrows represent CDSs. Red CDSs have an orthologue in the other genome, whereas blue CDSs do not. The vertical bars connect orthologous genes. Almost all of the genes present in the *L. johnsonii* gene cluster are also present in *L. plantarum*, but many of the genes in *L. plantarum* are not present in *L. johnsonii*. The genes unique to the *L. plantarum* cluster include genes encoding four cell-envelope proteins, three proteins of the pyruvate dehydrogenase complex, and a putative L-lactate dehydrogenase.

synteny could be detected with *L. lactis* or streptococci (data not shown).

Phylogenetic trees based on 16S RNA (Fig. 1) or highly conserved genes support the relatively large phylogenetic distance between *L. plantarum* and *L. johnsonii* suggested by the protein dot plot. An unrooted tree based on the *atpD* gene (part of the highly conserved ATP synthase cluster) shows *L. johnsonii* to be closely related to *L. gasseri*, but also shows a relatively large distance between *L. johnsonii* and *L. plantarum* (Siezen *et al.*, 2004), in agreement with Fig. 1. The phylogenetic distance between *L. plantarum* and *L. johnsonii* is in fact similar to the distance between *L. plantarum* and *E. faecalis*. These findings re-emphasize the difficulties in establishing the taxonomy of lactobacilli, and show that the current classification of the *Lactobacillus* genus, based on morphology and lactic acid production, is not always supported by phylogenetic relationships based on sequence homology and genome synteny.

Functional comparison of proteomes

The percentage of the total number of proteins of *L. plantarum* and *L. johnsonii* belonging to selected COG functional classes is shown in Fig. 4. Only classes displaying large differences between the two organisms are shown; Supplementary Table S2 shows the number of proteins of *L. plantarum* and *L. johnsonii* for all COG classes. This overview gives an indication of the differences in focus on metabolism and other cellular functions of these bacteria. Compared to *L. johnsonii*, *L. plantarum* has a relatively high number of proteins for carbohydrate, amino acid and lipid metabolism. Due to its smaller genome size, *L. johnsonii* has a higher percentage of genes involved in 'core functions' such as replication and translation.

Metabolic pathways

Metabolic reconstruction can provide insights into the differences and similarities in the metabolic potential of *L. plantarum* and *L. johnsonii*, which can be helpful in both explaining observed physiological differences between the two species and in the design of experimental studies to investigate genotype–phenotype relationships.

The mapping of enzymic functions on the metabolic pathways provided by the KEGG database resulted in the identification of a set of enzymes required for known biochemical pathways. The main differences between *L. johnsonii* and *L. plantarum* are listed in Supplementary Table S3. The classes and metabolic pathways that display striking differences between the two organisms will be described in some detail below.

The *L. plantarum* genome encodes 268 proteins predicted to be involved in the metabolism and transport of amino acids, while the *L. johnsonii* genomes encodes only 125. *L. plantarum* encodes the enzymes required for the biosynthesis of all amino acids, with the exception of leucine, isoleucine and valine. In contrast, *L. johnsonii* is predicted to be incapable of synthesizing most, if not all, of the 20 standard amino acids. This reflects the environmental niches in which the bacteria live. *L. johnsonii* typically is found only in the gut, although recent reports (Guan le *et al.*, 2003; Meroth *et al.*, 2003) suggest that *L. johnsonii* might also occur in other nutrient-rich environments, where it can take up amino acids and peptides from its environment. To this end, *L. johnsonii* has an extracellular, cell-bound proteinase to liberate these peptides from proteinaceous substrates, and more intracellular peptidases for degradation of imported peptides than *L. plantarum* (Kleerebezem *et al.*, 2003; Pridmore *et al.*, 2004). In contrast,

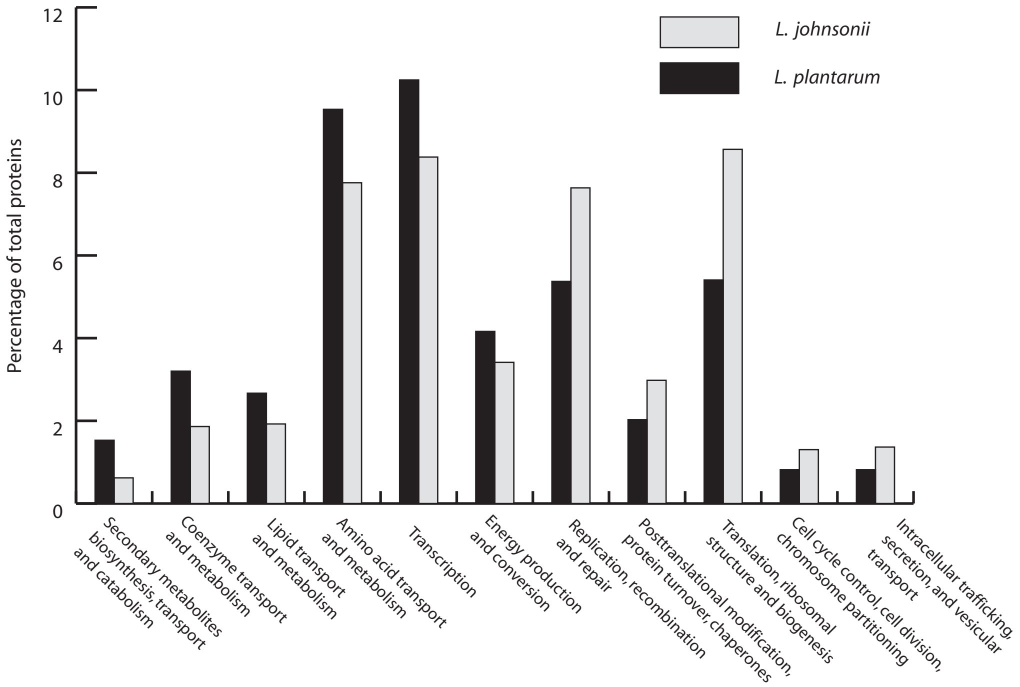

Fig. 4. COG classification of *L. johnsonii* and *L. plantarum* proteins. Only COG families displaying major differences between the two organisms are shown.

L. plantarum is also found in other environments, such as on plants and plant-derived materials, where amino acids and peptides are not as readily available, and hence has retained more amino acid biosynthetic capability.

While the *L. plantarum* genome encodes 90 proteins predicted to be involved in the transport and metabolism of vitamins and cofactors, the *L. johnsonii* genome encodes only 30. For instance, all the enzymes necessary for the biosynthesis of folate are present in *L. plantarum*. In contrast, *L. johnsonii* has only a few enzymes that could have a function in this pathway, but all of these enzymes could also have functions in other processes. This suggests that *L. plantarum* is capable of synthesizing its own folate, while *L. johnsonii* is not, which has recently been confirmed experimentally (Sybesma *et al.*, 2003).

Both *L. johnsonii* and *L. plantarum* have the capacity to synthesize pyrimidines *de novo*. However, only the *L. plantarum* genome encodes the proteins essential for *de novo* synthesis of purines from phosphoribosyl pyrophosphate; *L. johnsonii* needs inosine, which can be converted to IMP in a single enzymic step. This is consistent with the observation that *L. johnsonii* needs to obtain purines or their precursors from its environment (Elli *et al.*, 2000).

The *L. plantarum* genome encodes 13 proteins predicted to be involved in the biosynthesis of fatty acids, while the *L. johnsonii* genome encodes only one. However, the route by which *L. johnsonii* acquires fatty acids is still unknown.

L. plantarum can utilize a much wider variety of sugars than *L. johnsonii* (Supplementary Table S4). This corroborates the observation that many more proteins involved in the uptake, interconversion and degradation of sugars are encoded by the *L. plantarum* genome than by the *L. johnsonii* genome: the *L. plantarum* genome encodes 342 proteins of the COG class 'carbohydrate transport and metabolism', while the *L. johnsonii* genome encodes only 196.

L. plantarum has a more versatile pyruvate metabolism than *L. johnsonii* (Fig. 5). Both *L. plantarum* and *L. johnsonii* can convert pyruvate to L- and D-lactate, but *L. johnsonii* lacks the pyruvate dehydrogenase complex and other enzymes required for the conversion of pyruvate to acetate, acetaldehyde and acetyl-coenzyme A. Moreover, *L. plantarum* has a much higher redundancy of enzymes involved in pyruvate metabolism (Supplementary Table S5). *L. plantarum* is a facultative heterofermentative organism, capable of mixed-acid fermentation forming lactate, formate and/or acetate depending on environmental conditions, while *L. johnsonii* is an obligate homofermentative organism, capable of homolactic fermentation only. The lack of the pyruvate dehydrogenase complex in *L. johnsonii* is consistent with the anaerobic environment in the gastrointestinal tract

Cellular transport

Transporters enable uptake of essential nutrients, ions and metabolites, as well as the expulsion of toxic compounds,

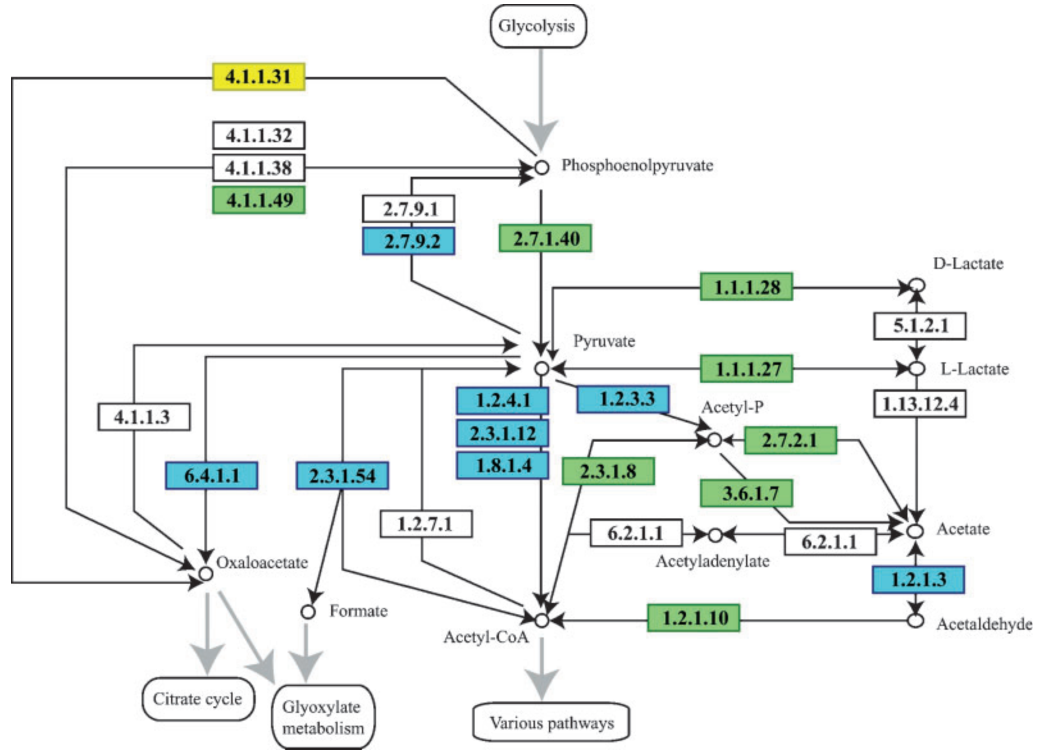

Fig. 5. Pyruvate metabolism in *L. johnsonii* and *L. plantarum*. The figure is based on the pyruvate metabolism pathway from the KEGG database (Kanehisa *et al*., 2002). Open circles represent metabolites; the square boxes represent enzymes with their EC numbers. The colour of a box indicates the presence of the gene encoding that enzyme in *L. plantarum* (blue), in *L. johnsonii* (yellow) or in both (green). Not shown are enzymes and pathways that (i) are cytochrome dependent, (ii) do not occur in bacteria, and (iii) have no known genes.

cell-envelope macromolecules and the end products of cellular metabolism. Putative transporters have been identified in *L. johnsonii* and *L. plantarum* by comparison with the Transport Classification DataBase (TC-DB; Busch & Saier, 2002) (Table 3). *L. johnsonii* contains 286 genes associated with various transport systems, accounting for more than 15 % of its total CDSs, which is proportionally slightly more than *L. plantarum* WCSF1 (473 proteins, 13 %). Both numbers compare well with other organisms of similar genome size living in nutrient-rich environments, such as the cheese starter *Lactococcus lactis* (11 % transporters; Bolotin *et al*., 2001) and the oral pathogen *Streptococcus mutans* (15 %; Ajdić *et al*., 2002).

The increased transport potential of *L. plantarum* is primarily due to an increased redundancy of transport

Table 3. Summary of transporters encoded in the genomes of *L. plantarum* and *L. johnsonii*

Transporter class	Number in:	
	L. plantarum	***L. johnsonii***
Channel proteins	22	10
Electrochemical potential-driven transporters	142	71
Primary active transporters*	195	148
Sugar-transporting phosphotransferase systems (PTS)	57	44
Transmembrane electron carriers	3	0
Accessory factors involved in transport	20	2
Incompletely characterized transport systems	34	11
Total	473	286

*Includes 147 and 105 ABC transporter proteins, respectively.

proteins. For instance, *L. plantarum* encodes six glycerol-uptake facilitator proteins, compared to a single protein in *L. johnsonii*. This observation suggests the importance of glycerol uptake in *L. plantarum*.

The most notable electrochemical potential-driven transporters in *L. johnsonii* are two conjugated bile salt–proton symporters (LJ0057 and LJ0058), which have been found to be unique proteins of the *Lactobacillus acidophilus* group of organisms (Pridmore *et al.*, 2004). Striking differences are found in the number of multidrug/oligosaccharidyl-lipid/polysaccharide flippase superfamily, the auxin-efflux carrier family and the drug/metabolite transporter superfamily of transporters in *L. plantarum*. The *L. plantarum* genome encodes 5, 4 and 11 proteins belonging to these families, respectively, whereas the *L. johnsonii* genome encodes only one protein of each family. Primary active transporters (mainly ABC transporters: 147 and 105 proteins in *L. plantarum* and *L. johnsonii*, respectively) represent the largest group of transporters in both lactobacilli. In *L. johnsonii* and *L. plantarum*, 16 and 25 complete PEP-dependent, phosphoryl transfer-driven group translocators (phosphotransferase system, PTS) systems were identified, respectively, including multiple systems for the uptake of glucose, mannose, fructose, and β-glucosides, and single systems for cellobiose, sucrose, and galactitol.

Extracellular proteins

Extracellular proteins are considered to be important for interaction of bacteria with their environment, for example in adhesion and communication. This makes them of special interest in the case of lactobacilli, because they may be involved in host–microbe and microbe–microbe interactions, such as in the gastrointestinal tract or on plant materials. Putative extracellular proteins of *L. plantarum* and *L. johnsonii* were identified by the presence of a Sec-pathway-dependent signal peptide. Both proteins that are secreted into the environment and proteins that become attached to the cell surface fall into this category. The latter were identified by searching for cell-anchoring domains, such as the N-terminal lipoprotein motif for anchoring to the cell membrane (Sutcliffe & Russell, 1995) and the C-terminal LPxTG motif for anchoring to peptidoglycan (Navarre & Schneewind, 1999). The *L. plantarum* and *L. johnsonii* genomes are predicted to encode 211 (Kleerebezem *et al.*, 2003) and 117 putative extracellular proteins, respectively. Nearly 90 % of these proteins in both species are predicted to contain at least one type of cell-wall anchoring domain.

A comparison of the putative extracellular proteins encoded in both genomes is summarized in Table 4. The set of extracellular proteins of known function is very similar in

Table 4. Comparison of putative extracellular proteins encoded in the genomes of *L. plantarum* and *L. johnsonii*

Functional Class	Number in:	
	L. plantarum	*L. johnsonii*
ABC transporters, substrate-binding domain	30	20
Regulators*	5	5
Enzymes, known functions		
Proteases	14	8
Transpeptidases	4	4
Cell-wall hydrolases	12	5
Other	5	6
Miscellaneous, known functions	8	4
Unknown function, present in both genomes		
Homologues, singles	19	20
Homologues, families†		
CSH1 family	8	5
CSH2 family	5	1
WY-domain family	4	2
Unknown function, present in single genome	97	37
Total	211	117

*Membrane-anchored proteins with extracellular transcriptional attenuator domain (Pfam PF03816).

†CSH1, cell-surface hydrolase family 1: contains active site triad Ser, Asp and His residues and consensus GxSHG typical of hydrolases (e.g. lipase, esterase). CSH2, cell-surface hydrolase family 2: contains active site triad Ser, Asp and His residues and consensus GxSMG typical of hydrolases (e.g. lipase, esterase). WY, C-terminal cell-surface binding domain (Jankovic *et al.*, 2003).

both lactobacilli, although *L. plantarum* has more paralogues for several of these known functions. However, the majority (55–65 %) of putative extracellular proteins are of unknown function (Table 4). Some of these are present in both lactobacilli, either as single copies of orthologues, or as multiple copies (paralogues) belonging to different families. Two families of putative cell-surface hydrolases (CSH-1 and CSH-2) are detected which have sequence characteristics of lipases or esterases (Anthonsen *et al.*, 1995; Wong & Schotz, 2002). It is striking to note that the majority of extracellular proteins with unknown function are not shared by *L. plantarum* and *L. johnsonii*, but only occur in one of the two bacteria.

The C-terminal LPxTG motif for covalent binding to peptidoglycan is present in 25 and 14 extracellular proteins of *L. plantarum* and *L. johnsonii*, respectively. Generally, these are large, multi-domain, repeat-containing proteins (Kleerebezem *et al.*, 2003; Pridmore *et al.*, 2004). Again, there is very little homology between the LPxTG proteins of *L. plantarum* and those of *L. johnsonii*, other than in the peptidoglycan attachment motif.

Regulators

Regulatory proteins play an important role in the adaptation of an organism to different environments. The *L. plantarum* genome is predicted to encode 264 regulators (9·4 % of all proteins), while the *L. johnsonii* genome has only 114 putative regulators (6 %), as summarized in Supplementary Table S6. This agrees with the general observation that large genomes have a relatively high number of proteins involved in transcription and regulation (Konstantinidis & Tiedje, 2004; van Nimwegen, 2003).

Besides the difference in genome size, the different lifestyles of *L. plantarum* and *L. johnsonii* also contribute to this difference. The number of proteins predicted to be involved in the regulation of sugar and energy metabolism is especially high in *L. plantarum*. This is in agreement with the differences found in sugar metabolism in the two organisms: *L. plantarum* can utilize a much wider variety of sugars than *L. johnsonii* (Supplementary Table S4). *L. plantarum* with its free-living lifestyle needs to be capable of dealing with many different environmental circumstances (Boneca *et al.*, 2003), and apparently has both the metabolic capacity and the regulatory machinery to deal with adaptation to different niches, while *L. johnsonii* does not need a complex regulatory apparatus because of the relatively stable environment in the gastrointestinal tract.

LAB-specific and unique genes

A Smith & Waterman homology search was used to identify proteins unique to either *L. plantarum* or *L. johnsonii*, and proteins unique to LAB (Table 5). The table lists the number of proteins that are present in either *L. plantarum* or *L. johnsonii* and in at least one other LAB, but without homologues in organisms not considered as LAB. It also lists the number of proteins found to be unique to either *L. plantarum* or *L. johnsonii*. The individual proteins for these categories can be found in Supplementary Tables S7–S12. The result of this analysis depends of course on the number of genomes available at the time of comparison, and is only preliminary, since many of the LAB genomes in the ERGO database were less than 100 % complete at the time of this analysis.

We identified 181 and 243 genes in the *L. plantarum* and *L. johnsonii* genomes, respectively, that encode proteins with homologues only in other LAB. Of those, only about 40 proteins are shared between *L. plantarum* and *L. johnsonii* (Table 5). In contrast, the *L. plantarum* genome encodes 143 proteins with homologues only in other LAB, but without a homologue in *L. johnsonii*. This number is much

Table 5. Unique and LAB-specific proteins in *L. plantarum* and *L. johnsonii*

Proteins	Number in:	
	L. plantarum	***L. johnsonii***
Total number of proteins	3009	1821
Proteins with homologues in non-LAB	2407	1466
Homologues in both *L. plantarum* and *L. johnsonii*	1549	1256
Homologues in *L. plantarum*, but not in *L. johnsonii*	858	–
Homologues in *L. johnsonii*, but not in *L. plantarum*	–	210
Proteins with homologues only in other LAB*	181	243
Homologues in both *L. plantarum* and *L. johnsonii*	38	47
Homologues in *L. plantarum*, but not in *L. johnsonii*	143	–
Homologues in *L. johnsonii*, but not in *L. plantarum*	–	196
Genome-specific proteins†	421	112

*See Supplementary Tables S7, S8, S10 and S11 for details.

†See Supplementary Tables S9 and S12 for details.

lower than the 196 LAB-specific proteins encoded by *L. johnsonii* without homologues in *L. plantarum*, especially considering the relatively large size of the *L. plantarum* genome compared to the *L. johnsonii* genome. This difference is caused by the close relatedness of the *L. johnsonii* and *L. gasseri* genomes; these two organisms share the same niche and have a very similar genetic make-up and genome organization (Pridmore *et al.*, 2004). Moreover, this also explains the relatively low number of unique genes in *L. johnsonii*.

Many of the proteins present in *L. plantarum* but absent in *L. johnsonii*, or vice versa, are grouped in clusters on the genome. A large number of these clustered unique genes encode functionally related proteins, such as those involved in the biosynthesis of polysaccharides, bacteriocins and prophages (Supplementary Table S13). In *L. plantarum*, such clusters frequently have a high base-deviation index (BDI), suggesting horizontal transfer (Kleerebezem *et al.*, 2003). In *L. johnsonii* however, only the polysaccharide biosynthesis cluster (LJ1027–1047) has a high BDI.

Most of the proteins predicted to be LAB specific are of unknown function (Supplementary Tables S7–S12). The identification of structural features, such as signal peptides, transmembrane helices and cell-wall anchors, and conserved domains/motifs in these proteins, such as those involved in the binding of ATP, DNA and carbohydrates, could be used to predict their function and to identify potentially interesting targets for future research. In this way, the preliminary analysis of LAB-specific genes described here can serve as a starting point for a more comprehensive study of LAB-specific proteins and gene clusters, once the complete genome sequences of many LAB species become available (Klaenhammer *et al.*, 2002).

Concluding remarks

The ability of *L. plantarum* to survive in many different environments is reflected by the much more elaborate metabolic, regulatory and transport machinery compared to that of *L. johnsonii*. The differences between *L. plantarum* and *L. johnsonii*, both in genome organization and in gene content, are exceptionally large for two bacteria of the same genus (Suyama & Bork, 2001). Similar differences have been reported only in streptococci (Tettelin *et al.*, 2002). This low degree of synteny between *L. plantarum* and *L. johnsonii* suggests that they are only marginally more related to each other than to other Gram-positive bacteria. These findings emphasize the difficulty in taxonomic classification of lactobacilli.

Overall, the genome-wide comparison of two complete *Lactobacillus* genomes has provided unique information on the relatedness and differences between the two species. This has led to insight into the genomic adaptation to ecological niches of *L. plantarum* and *L. johnsonii*, and provides leads for targeted experimental studies.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Netherlands Organization for Scientific Research, (NWO-BMI project 050.50.206). We thank Douwe Molenaar for the calculation of BDI, Quinta Helmer for the analysis of extracellular proteins and John Janssen for the identification of conserved gene clusters. We thank Dr T. Klaenhammer (North Carolina State University) for access to the incomplete genome of *Lactobacillus gasseri* and Dr C. Tran and Dr M. Saier (University of California in San Diego) for making TC-DB available.

REFERENCES

Ajdić, D., McShan, W. M., McLaughlin, R. E. & 16 other authors (2002). Genome sequence of *Streptococcus mutans* UA159, a cariogenic dental pathogen. *Proc Natl Acad Sci U S A* **99**, 14434–14439.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403–410.

Anthonsen, H. W., Baptista, A., Drablos, F., Martel, P., Petersen, S. B., Sebastiao, M. & Vaz, L. (1995). Lipases and esterases: a review of their sequences, structure and evolution. *Biotechnol Annu Rev* **1**, 315–371.

Boeckmann, B., Bairoch, A., Apweiler, R. & 9 other authors (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365–370.

Bolotin, A., Wincker, P., Mauger, S., Jaillon, O., Malarme, K., Weissenbach, J., Ehrlich, S. D. & Sorokin, A. (2001). The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. lactis IL1403. *Genome Res* **11**, 731–753.

Boneca, I. G., de Reuse, H., Epinat, J. C., Pupin, M., Labigne, A. & Moszer, I. (2003). A revised annotation and comparative analysis of *Helicobacter pylori* genomes. *Nucleic Acids Res* **31**, 1704–1714.

Braun-Fahrlander, C., Riedler, J., Herz, U. & 12 other authors (2002). Environmental exposure to endotoxin and its relation to asthma in school-age children. *N Engl J Med* **347**, 869–877.

Busch, W. & Saier, M. H., Jr (2002). The transporter classification (TC) system, 2002. *Crit Rev Biochem Mol Biol* **37**, 287–337.

Dandekar, T., Snel, B., Huynen, M. & Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **23**, 324–328.

Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**, 4636–4641.

Eddy, S. R. (1996). Hidden Markov models. *Curr Opin Struct Biol* **6**, 361–365.

Elli, M., Zink, R., Rytz, A., Reniero, R. & Morelli, L. (2000). Iron requirement of *Lactobacillus* spp. in completely chemically defined growth media. *J Appl Microbiol* **88**, 695–703.

Fujisawa, T., Benno, Y., Yaeshima, T. & Mitsuoka, T. (1992). Taxonomic study of the *Lactobacillus acidophilus* group, with recognition of *Lactobacillus gallinarum* sp. nov. and *Lactobacillus johnsonii* sp. nov. and synonymy of *Lactobacillus acidophilus* group A3 (Johnson *et al.*, 1980) with the type strain of *Lactobacillus amylovorus* (Nakamura 1981). *Int J Syst Bacteriol* **42**, 487–491.

Guan le, L., Hagen, K. E., Tannock, G. W., Korver, D. R., Fasenko, G. M. & Allison, G. E. (2003). Detection and identification of *Lactobacillus* species in crops of broilers of different ages by using PCR-denaturing gradient gel electrophoresis and amplified ribosomal DNA restriction analysis. *Appl Environ Microbiol* **69**, 6750–6757.

Haller, D., Blum, S., Bode, C., Hammes, W. P. & Schiffrin, E. J. (2000a). Activation of human peripheral blood mononuclear cells by nonpathogenic bacteria in vitro: evidence of NK cells as primary targets. *Infect Immun* **68**, 752–759.

Haller, D., Bode, C., Hammes, W. P., Pfeifer, A. M., Schiffrin, E. J. & Blum, S. (2000b). Non-pathogenic bacteria elicit a differential cytokine response by intestinal epithelial cell/leucocyte co-cultures. *Gut* **47**, 79–87.

Ibnou-Zekri, N., Blum, S., Schiffrin, E. J. & von der Weid, T. (2003). Divergent patterns of colonization and immune response elicited from two intestinal *Lactobacillus* strains that display similar properties in vitro. *Infect Immun* **71**, 428–436.

Jankovic, I., Ventura, M., Meylan, V., Rouvet, M., Elli, M. & Zink, R. (2003). Contribution of aggregation-promoting factor to maintenance of cell shape in *Lactobacillus gasseri* 4B2. *J Bacteriol* **185**, 3288–3296.

Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. (2002). The KEGG databases at GenomeNet. *Nucleic Acids Res* **30**, 42–46.

Klaenhammer, T., Altermann, E., Arigoni, F. & 33 other authors (2002). Discovering lactic acid bacteria by genomics. *Antonie Van Leeuwenhoek* **82**, 29–58.

Kleerebezem, M., Boekhorst, J., van Kranenburg, R. & 17 other authors (2003). Complete genome sequence of *Lactobacillus plantarum* WCFS1. *Proc Natl Acad Sci U S A* **100**, 1990–1995.

Klein, G. (2003). Taxonomy, ecology and antibiotic resistance of enterococci from food and the gastro-intestinal tract. *Int J Food Microbiol* **88**, 123–131.

Konstantinidis, K. T. & Tiedje, J. M. (2004). Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc Natl Acad Sci U S A* **101**, 3160–3165.

Link-Amster, H., Rochat, F., Saudan, K. Y., Mignot, O. & Aeschlimann, J. M. (1994). Modulation of a specific humoral immune response and changes in intestinal flora mediated through fermented milk intake. *FEMS Immunol Med Microbiol* **10**, 55–63.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting protein function and protein–protein interactions from genome sequences. *Science* **285**, 751–753.

Meroth, C. B., Walter, J., Hertel, C., Brandt, M. J. & Hammes, W. P. (2003). Monitoring the bacterial population dynamics in sourdough fermentation processes by using PCR-denaturing gradient gel electrophoresis. *Appl Environ Microbiol* **69**, 475–482.

Metzger, R., Brown, D. P., Grealish, P., Staver, M. J., Versalovic, J., Lupski, J. R. & Katz, L. (1994). Characterization of the macromolecular synthesis (MMS) operon from *Listeria monocytogenes*. *Gene* **151**, 161–166.

Navarre, W. W. & Schneewind, O. (1999). Surface proteins of Gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. *Microbiol Mol Biol Rev* **63**, 174–229.

Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int J Neural Syst* **8**, 581–599.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96**, 2896–2901.

Page, R. D. (1996). TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* **12**, 357–358.

Pridmore, D., Berger, B., Desiere, F. & 12 other authors (2004). The genome sequence of the probiotic intestinal bacterium *Lactobacillus johnsonii* NCC 533. *Proc Natl Acad Sci U S A* **101**, 2512–2517.

Schiex, T., Gouzy, J., Moisan, A. & de Oliveira, Y. (2003). FrameD: a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. *Nucleic Acids Res* **31**, 3738–3741.

Shazand, K., Tucker, J., Grunberg-Manago, M., Rabinowitz, J. C. & Leighton, T. (1993). Similar organization of the nusA-infB operon in *Bacillus subtilis* and *Escherichia coli*. *J Bacteriol* **175**, 2880–2887.

Siezen, R. J., Van Enckevort, F. H., Kleerebezem, M. & Teusink, B. (2004). Genome data mining of lactic acid bacteria: the impact of bioinformatics. *Curr Opin Biotechnol* **15**, 105–115.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol* **147**, 195–197.

Snel, B., Bork, P. & Huynen, M. A. (2002). The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci U S A* **99**, 5890–5895.

Sonnhammer, E. L. & Durbin, R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**, GC1–GC10.

Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A. & Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* **26**, 320–322.

Sutcliffe, I. C. & Russell, R. R. (1995). Lipoproteins of Gram-positive bacteria. *J Bacteriol* **177**, 1123–1128.

Suyama, M. & Bork, P. (2001). Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends Genet* **17**, 10–13.

Sybesma, W., Starrenburg, M., Tijsseling, L., Hoefnagel, M. H. & Hugenholtz, J. (2003). Effects of cultivation conditions on folate production by lactic acid bacteria. *Appl Environ Microbiol* **69**, 4542–4548.

Tamames, J. (2001). Evolution of gene order conservation in prokaryotes. *Genome Biol* **2**, research0020.1–0020.11.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V. & 7 other authors (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29**, 22–28.

Tettelin, H., Nelson, K. E., Paulsen, I. T. & 36 other authors (2001). Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* **293**, 498–506.

Tettelin, H., Masignani, V., Cieslewicz, M. J. & 40 other authors (2002). Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. *Proc Natl Acad Sci U S A* **99**, 12391–12396.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680.

Tillier, E. R. & Collins, R. A. (2000). Genome rearrangement by replication-directed translocation. *Nat Genet* **26**, 195–197.

van Nimwegen, E. (2003). Scaling laws in the functional content of genomes. *Trends Genet* **19**, 479–484.

Vaughan, E. E., de Vries, M. C., Zoetendal, E. G., Ben-Amor, K., Akkermans, A. D. & de Vos, W. M. (2002). The intestinal LABs. *Antonie Van Leeuwenhoek* **82**, 341–352.

Wong, H. & Schotz, M. C. (2002). The lipase gene family. *J Lipid Res* **43**, 993–999.

Wuyts, J., Perriere, G. & Van De Peer, Y. (2004). The European ribosomal RNA database. *Nucleic Acids Res* **32**, Database issue D101–D103.