The genome sequence of the probiotic intestinal bacterium *Lactobacillus johnsonii* NCC 533

R. David Pridmore*†, Bernard Berger*, Frank Desiere*, David Vilanova*, Caroline Barretto*, Anne-Cecile Pittet*, Marie-Camille Zwahlen*, Martine Rouvet*, Eric Altermann‡, Rodolphe Barrangou‡, Beat Mollet*, Annick Mercenier*, Todd Klaenhammer‡, Fabrizio Arigoni*, and Mark A. Schell*§

*Department of Nutrition and Health, Nestlé Research Center, P.O. Box 44, Vers-chez-les-Blanc, 1000 Lausanne 26, Switzerland; ‡Department of Food Science, Southeast Dairy Foods Research Center, North Carolina State University, Raleigh, NC 27695-7624; and §Department of Microbiology, University of Georgia, Athens, GA 30602

Contributed by Todd Klaenhammer, December 29, 2003

Lactobacillus johnsonii **NCC 533 is a member of the acidophilus group of intestinal lactobacilli that has been extensively studied for their "probiotic" activities that include, pathogen inhibition, epithelial cell attachment, and immunomodulation. To gain insight into its physiology and identify genes potentially involved in interactions with the host, we sequenced and analyzed the 1.99-Mb genome of *L. johnsonii* NCC 533. Strikingly, the organism completely lacked genes encoding biosynthetic pathways for amino acids, purine nucleotides, and most cofactors. In apparent compensation, a remarkable number of uncommon and often duplicated amino acid permeases, peptidases, and phosphotransferase-type transporters were discovered, suggesting a strong dependency of NCC 533 on the host or other intestinal microbes to provide simple monomeric nutrients. Genome analysis also predicted an abundance (>12) of large and unusual cell-surface proteins, including fimbrial subunits, which may be involved in adhesion to glycoproteins or other components of mucin, a characteristic expected to affect persistence in the gastrointestinal tract (GIT). Three bile salt hydrolases and two bile acid transporters, proteins apparently critical for GIT survival, were also detected. *In silico* genome comparisons with the >95% complete genome sequence of the closely related *Lactobacillus gasseri* revealed extensive synteny punctuated by clear-cut insertions or deletions of single genes or operons. Many of these regions of difference appear to encode metabolic or structural components that could affect the organisms competitiveness or interactions with the GIT ecosystem.**

The human gastrointestinal tract (GIT) is a nutrient-rich environment that is colonized by a vast and complex collection of microorganisms that play a major role in its function and development (1, 2). The GIT microbiota consists of >500 species of bacteria, most of which remain uncultured, and whose composition varies with the individual, age, and location in the GIT (3). GIT microbes are active partners in polysaccharide and protein digestion, and ultimately are responsible for a major part of the GIT metabolic activity. They also produce vitamins, short-chain fatty acids, and other nutrients for their hosts, providing up to 15% of total caloric intake (3). It has been proposed that a balanced and diverse microbiota is essential for healthy intestinal function, as well as resistance to infection by enteric pathogens (4). As a result, several lactobacilli, a few bifidobacteria, and their fermented food products are extensively marketed as probiotic foods (5, 6).

Lactobacilli are nutritionally fastidious anaerobes in the low-GC Gram-positive group of the Lactobacillales that also includes Streptococcaceae, Enterococcaceae, and Leuconostocaceae (www.ncbi.nlm.nih.gov/Taxonomy). For energy metabolism they rapidly ferment sugars to lactic acid and have been historically and economically important in the fermentation and preservation of milk, vegetables, cereal, and meat products. Using non-culture-based methods, Marteau *et al.* (7) reported that the *Lactobacillus–Enterococcus* group represents 6% of the fecal microbiota and 23% of the microbiota in the cecum at the junction of the ileum and colon. Of the >50 known species of lactobacilli, the "acidophilus complex," composed of six closely related species, has received particular attention because of their reported probiotic properties (6) and prominence among bacteria found in human feces and vagina. A subgroup of the complex includes *Lactobacillus johnsonii* and *Lactobacillus gasseri* (8), with *L. gasseri*-related bacteria being reported as the dominant bacteria in human intestinal wall biopsy samples (9).

L. johnsonii NCC 533 (formerly *Lactobacillus acidophilus* La1), a human isolate (10), has been extensively studied for its probiotic-associated activities, including immunomodulation (11–13), pathogen inhibition (14), and epithelial cell attachment (15, 16). To rapidly advance our understanding of *Lactobacillus* physiology and to identify potential bacterial components involved in host interactions, we sequenced and analyzed the genome of *L. johnsonii* NCC 533. The genome sequence revealed an unexpected number of genes that are not widely distributed among prokaryotes and hence may be important for the ability of *L. johnsonii* NCC 533 to persist and compete in the complex ecosystem of the GIT.

Materials and Methods

The *L. johnsonii* NCC 533 genome was shotgun-sequenced to 12.7-fold coverage, assembled into 134 contigs with PHRED (17, 18), and manually edited. Gaps were closed by sequencing multiplex PCR products and overlapping BAC clones giving one contig of 1.99 Mb with an error rate of <1 per 10^5 nucleotides (deposited in GenBank under accession no. AE017198). ORFs were identified with FrameD (19), and their starts were adjusted manually based on plausible ribosome-binding sites and/or BLAST alignments. Intergenic regions were reanalyzed for ORFs by using TBLASTX (20). Predicted ORFs of >75 residues were compared to public databases by using BLASTP and BLASTX; motif analysis was performed by using HMMER on PFAM 5.4 (http://pfam.wustl.edu/textsearch.shtml) (21). Predicted ORFs were assigned to a NCBI COG family by Smith–Waterman comparison (22). After review of bioinformatic data, a function or description was manually assigned to each ORF. tRNAs were identified by using TRNASCAN-SE (23) with stringent parameters and the prokaryotic covariance model. Transporters were identified by using BLAST and the Transport Protein Database (http://tcdb.ucsd.edu/tcdb/background.php), with an *E*-value cutoff of $<10^{-4}$ (24). Bioinformatic comparisons of the *L. johnsonii* genome with *L. gasseri* ATCC 33323, *Lactobacillus*

Abbreviations: BSH, bile salt hydrolase; EPS, exopolysaccharide; GIT, gastrointestinal tract; LAB, lactic acid bacteria; PTS, phosphotransferase.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. AE017198).

†To whom correspondence should be addressed. E-mail: raymond-david.pridmore@rdls.nestle.com.

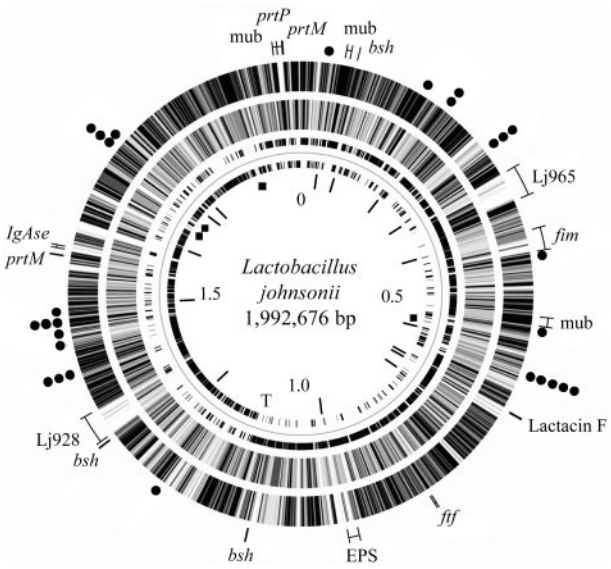

Fig. 1. Schematic of the *L. johnsonii* NCC 533 genome. The outer ring shows results of BLASTP analysis of *L. johnsonii* ORFs against *L. gasseri* ORFs; the darkness of the ORF lines is inversely proportional to the observed *E* value (i.e., darker denotes lower *E* value, more conserved; white, no homologs detected). The next ring shows results of a similar analysis of *L. johnsonii* ORFs against the nonredundant protein database. Ring 3 shows the position of the predicted genes on the leading strand (outer circle) and lagging strands (inner circle). Ring 4 shows positions of rRNA operons (squares), the 14 complete IS elements (lines), and the predicted replication terminus (T). Positions of peptidases (circles) and other genes discussed in the text (mub, mucus-binding protein; *prtP*, cell-wall protease; *prtM*, maturase; *fim*, fimbrae-like operon; *IgAse*, IgA protease; Lj928 and Lj965, prophages; *bsh*, bile salt hydrolase; *ftf*, fructosyltransferase) are indicated on the outside. The map was created by using GAMOLA (58) and GENEWIZ (59).

delbrueckii subsp. *bulgaricus*, *Lactococcus lactis*, *Lactobacillus breve*, *Lactococcus cremoris*, and *Lactobacillus casei* were done by using data from the Joint Genome Institute (http://spider.jgi-psf.org/JGI_microbial/html) and the Oak Ridge National Laboratory (http://compbio.ornl.gov/channel).

Results

General Genome Characteristics. The 1,992,676-bp genome of *L. johnsonii* NCC 533 is 34.6% G+C and contains six *rrn* operons at four loci, 79 tRNAs, and two complete prophages (25) (Fig. 1). We defined 1,821 coding regions, of which 77% (1,396) were attributed to a COG family and 75% (1,321) were given a functional description. Total GC-skew analysis and ORF direction reversal identified a potential origin of replication (OriC) as expected next to *dnaA* and a terminus at ≈1.1 Mb. We identified 14 complete IS (insertion sequence) elements from three known families: seven copies of IS*Ljo1*, an IS*30* family element (http://www-is.biotoul.fr/is.html); five intact copies of IS*Ljo2*, an IS*L3* family element; and one copy each of IS*Ljo4* and IS*Ljo5*, IS*200*–IS*650* family elements.

Predicted Biosynthetic Capabilities. *L. johnsonii* appears incapable of *de novo* synthesis of most, if not all, amino acids, because no complete or partial amino acid biosynthetic pathways were identified from the predicted ORFs. However, *L. johnsonii* may be able to synthesize Gly from Ser via serine hydroxymethyltransferase (*glyA*, LJ0263) and Asn from Asp by using asparagine synthase (*asnA*, LJ0511). It may also synthesize Asp from oxaloacetate and Glu, using aspartate aminotransferase (*aspC*, LJ1390). Additionally, the presence of glutamine synthetase (*glnA*, LJ1614) and a probable glutaminase homolog (*gls1*, LJ0713) indicates that NCC 533 is able to interconvert Glu and Gln. Thus, from Glu it can make Gln, Asp, and Asn. Nonetheless, it appears that *L. johnsonii* depends on large amounts of exogenous amino acids and/or peptides to fuel protein synthesis and hence will be restricted to environments that are rich in such substrates.

L. johnsonii does not appear to assimilate ammonium because it lacks homologs of glutamate dehydrogenase and glutamate synthetase. Likewise, no sulfur assimilation pathways were detected. Similar to *L. delbrueckii* and *L. gasseri*, *L. johnsonii* has phosphoenolpyruvate carboxylase and a partial TCA cycle with fumarate reductase (*fccA*, LJ1404), fumarate hydratase (*fumH*, LJ1405), and a possible malate dehydrogenase (LJ1742), presumably to produce oxaloacetate for aspartate synthesis and perhaps to dispose of excess reducing equivalents (NADH) generated during fermentation. *L. johnsonii* lacks homologs of the enzymes required to synthesize many cofactors such as thiamin, nicotinate, riboflavin, biotin, cobalamin, pantothenate, and pyridoxine.

L. johnsonii NCC 533 appears to contain all necessary genes for *de novo* synthesis of the pyrimidines dTMP, UMP, and CMP. In contrast, 8 of the 10 initial genes needed for *de novo* purine biosynthesis are absent. However, genes for conversion of inosine, xanthine, and hypoxanthine to IMP, GMP, and AMP are present. This finding is consistent with the observation that NCC 533 has four predicted proteins in the hypoxanthine or xanthine permease families and does not grow in defined media without inosine (H. van der Kaaij, personal communication). Thus, similar to the situation for amino acids, *L. johnsonii* is auxotrophic for purine nucleotides and must obtain them or their precursors from its environment. The auxotrophy of NCC 533 for nucleotides, cofactors, and amino acids predicted from the genome sequence is consistent with its known fastidious growth requirements (26).

Energy Metabolism and Transport. The genome sequence predicts *L. johnsonii* has a strict anaerobic energy metabolism fermenting a variety of disaccharides and hexoses to lactic acid, and perhaps some acetate via its Embden–Meyerhoff pathway. These sugars include some not previously recognized as *L. johnsonii* growth substrates: galactose, maltose, sorbose/sorbitol, gentiobiose, isoprimerevose, isomaltose, and panose (27). In contrast to *L. gasseri*, NCC 533 has acetolactate synthase and α-acetolactate decarboxylase (*alsS*, LJ1124 and *aldB*, LJ1125, respectively), so it likely converts some pyruvate to acetoin. It also may ferment ribose and, possibly, arabinose and xylose by using its aldose epimerase (*galM*, LJ0861) and xylulose-5-phosphate phosphoketolase (*xpkA*, LJ0803).

Because of its limited biosynthetic capabilities, *L. johnsonii* would be predicted to compensate by depending on enhanced transport capabilities to acquire cofactors, amino acids, and other essential precursors exogenously. Consistent with this, the codon adaptation index (28) of transporter genes suggest they are among the more highly expressed genes of NCC 533. Although the percentage of ORFs involved in transport (15%) is not too much higher than in other lactobacilli, the overabundance of certain types of less common transporters is notable. *L. johnsonii* has >20 AA-permease type transporters (PFAM 00324), more than twice the number found in most other lactic acid bacteria (LAB). Its genome also encodes 16 phosphotransferase (PTS)-type transporter systems, many more than nearly all other microbes with similar sized genomes, including *Enterococcus faecalis*; only *Listeria monocytogenes* has more. Although, the actual substrates of these 16 PTS transporters cannot be unambiguously predicted, they appear to transport many different sugars such as mannose, raffinose, *N*-acetyl glucosamine, trehalose, cellobiose, melibiose, and sucrose. These predictions are well supported by the *L. johnsonii* sugar fermentation profile

described by Fujisawa *et al.* (27) and its API 50CH profile (unpublished results). Fermentation substrates of *L. johnsonii* appear to be largely restricted to mono-, di-, and trisaccharides, because it lacks xylanases, amylases, arabinofuranosidases, and some other enzymes that depolymerize higher order complex polysaccharides. However, the finding of a putative neopullulanase (LJ0212) implies that NCC 533 may be able to use pullulan, a polymer of α 1–6 linked maltotriose that is an intermediate in the breakdown of branched starches.

Regulation. The numerically predominant regulatory protein families in NCC 533 are negative regulators: GntR (nine members), LacI (seven members), RpiR (five members), and ArsR (three members). Many of these are likely to be involved in regulation of sugar metabolism because they are located in gene clusters predicted to be involved in sugar acquisition and utilization. Relative to its genome size, NCC 533 has more predicted RpiR phospho-sugar responsive repressors than most other bacterial genomes, with the exception of *Enterococcus* spp. A relatively high number of RpiR and GntR regulators appears to be a general feature of the *Lactobacillus* genomes characterized to date. This finding is consistent with the extensive reliance of the LAB on PTS sugar transporters for import of phospho sugars as the major substrates for fermentation.

The major transcriptional activator families represented in the NCC 533 genome were LysR and AraC (four members each) and two-component systems (nine members). NCC 533 has no additional sigma factors other than the primary sigma factor RpoD. Homologs of the global regulators HrcA, LuxS, CcpA, and LexA, as well as the transcription termination regulators NusB, NusG, and GreAB are also present. A similar content and distribution of regulators was found in the genomes of *L. gasseri*, *Lactococcus cremoris*, and *L. casei*. Homologs of the global regulator Fur (ferric uptake regulator) are absent from genomes of *L. johnsonii* and some other lactobacilli (*L. brevis*, *L. casei*, *L gasseri*, and *L. delbrueckii*), implying that iron availability is not a major physiological concern for these bacteria. The apparent simplicity of the transcriptional regulatory networks in NCC 533 may reflect the richness and relative stability of the GIT environment.

Proteases and Peptidases. Because NCC 533 is totally dependent on exogenous amino acids for growth, it is not surprising that it has an unusually large number and types of proteinases, peptide transporters, and peptidases to obtain these from proteinaceous substrates in the GIT. NCC 533 is predicted to have an extracellular, cell-wall bound proteinase (LJ1840), three oligopeptide transporters (one classic ABC-type and two di- and tripeptide types), and an extensive collection of >25 cytoplasmic peptidases (Table 1, which is published as supporting information on the PNAS web site). Phylogenetic analysis of selected peptidases from NCC 533 and some other LAB generally showed closer relationships between the *Lactobacillus plantarum* and *Lactococcus* peptidases. These were more distant to the *L. johnsonii–L. gasseri* group, which were closely related to each other. In contrast, for proline peptidases, the *L. plantarum* enzymes aligned closer to those of other intestinal lactobacilli and were more distant from *Lactococcus* enzymes.

Aminopeptidases and dipeptidases (six each predicted in the C1-like and U34 families; PF03051 and PF03577, respectively) are particularly abundant. No other *Lactobacillus* sp. has more C1-like peptidases than NCC 533, possibly because some have arisen by duplication. This is evidenced by two pairs of tandem aminopeptidases (LJ0176/0178 and LJ0716/0719), which show ≈60% amino acid sequence identity to each other. Interestingly, homologs of LJ0176/0178 are limited to genomes of a few closely related GIT-inhabiting lactobacilli, some streptococci, and surprisingly, a few phylogenetically distant GIT inhabitants such as *Bifidobacterium longum* and *Bacteroides fragilis* suggesting the specificity of these may reflect on the structure and abundance of peptides present in the GIT. *L. johnsonii* NCC 533 also has more U34-type dipeptidases than any other *Lactobacillus* spp., again with the evidence suggesting that some of these pairs (e.g., LJ0545/0745 and LJ0653/0718) arose by gene duplication. Interestingly, many peptidases are clustered in possible operons with amino acid/peptide transporters and other genes seemingly involved in peptide metabolism. An example is the LJ0713–LJ0727 cluster containing a glutaminase, two very similar aminopeptidases, two very similar (>75% identity) PotE-type glutamate antiporters, a dipeptidase, a cationic amino acid permease, and a gutaminyl amino peptidase.

LJ1840 encodes a 2,209-residue protein with 70% similarity to the cell envelope-associated proteinase PrtH of *Lactobacillus helveticus* (29). Strong homologs are detected only in the milk-fermenting bacteria, *Lactococcus lactis*, *L. delbrueckii*, *L. helveticus*, and *Lactobacillus paracasei*. In these bacteria, this enzyme is critical for degradation of milk casein to oligopeptides that then serve as source of amino acids (30). Although the role of this protease in NCC 533 is unclear, it may enhance acquisition of amino acids in the competitive GIT environment or participate in host interactions via proteolytic cleavage of glycoproteins in the mucus layer (31).

It is clear that NCC 533 is programmed for amino acid and peptide transport and utilization from an environment where these must be readily available. In contrast, *L. plantarum* and *Lactococcus lactis* harbor fewer peptidases but encode more amino acid biosynthetic capability that supports their documented ability to grow and survive in environments deficient in amino acids.

Predicted Adhesion Factors. Cell-surface proteins and polysaccharides play important roles in microbial attachment to mucosal surfaces (16, 32). Presumably they assist in transient colonization of the GIT surface but also could actively participate in interactions with the intestinal epithelial and immune cells, thereby providing a route to immune modulation. Analysis of NCC 533 for proteins with signal peptides by using SIGNALP (33) and PROSITE motif PS00013 identified 42 potential cell-surface lipoproteins (Table 2, which is published as supporting information on the PNAS web site). Sixteen of these appeared to encode solute-binding proteins associated with ABC transporters. Other noteworthy predicted lipoproteins possibly involved in host–bacteria interactions are LJ1816, having 38% identity to the $CD4^+$ T cell-stimulating antigen of *L. monocytogenes* (34), and LJ0577, having 25% identity to the saliva-binding protein from *Streptococcus sanguis* (35).

Fourteen large ORFs were identified with both a signal sequence and a Gram-positive cell-wall anchor motif (PFAM PF00746), implying that they are secreted and attached to the cell surface (Table 3, which is published as supporting information on the PNAS web site). Most of these ORFs are composed >1,000 residues, and like many other cell-surface proteins (36) have regions with highly repetitive sequences. BLAST analysis with other *Lactobacillus* genomes implied that LJ0382, LJ0391, LJ1128, LJ1711, and LJ1839 are specific to *L. johnsonii* NCC 533. However, genomes of both *L. gasseri* and *L. plantarum* appear to encode several proteins with analogous size, organization, and, to a lesser extent, amino acid sequence. LJ0047, LJ0484, and LJ1839 are of special interest because they share significant similarity ($E < 10^{-44}$) to the mucus-binding protein (MUB) of *Lactobacillus reuteri* (32), a 358-kDa surface protein that appears to specifically bind to mucin glycoproteins. Thus, these MUB homologs may be involved in the reported protease-sensitive ability of NCC 533 to bind intestinal mucins *in vitro* (37). Analogous to the amino acid repeats found in *L. reuteri* MUB, the C-terminal end of LJ0047 contains four repeats with the

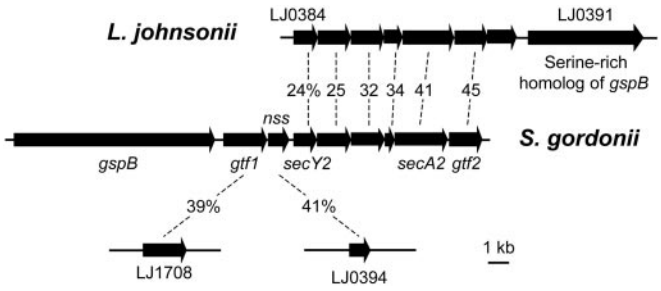

Fig. 2. Comparison of the *S. gordoni gspB* fimbrial operon (GenBank accession no. AY028381) to the homologous segment of *L. johnsonii* NCC 533. LJ1708 and LJ0394, which are located elsewhere in the NCC 533 genome, are also homologous counterparts in the *S. gordoni gspB* fimbrial operon. Numbers adjacent to dotted lines show amino acid sequence identity between homologs.

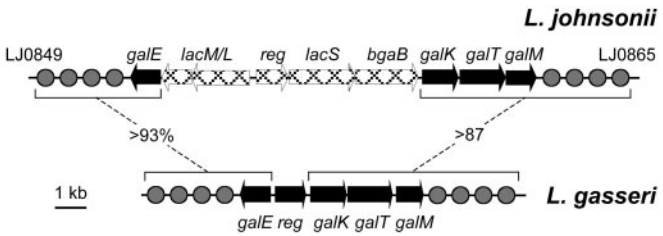

Fig. 3. Alignment of genomic regions of *L. gasseri* and *L. johnsonii* NCC 533 containing the galactose utilization genes. Hatched arrows represent lactose catabolism genes of NCC 533 that were apparently inserted into the common *L. gasseri–L. johnsonii* backbone. Homologous and syntenic genes flanking the insertion are indicated by gray circles and black arrows (*gal* genes). *reg* denotes a *lacI*-type regulator gene. The minimum percentage identity between the group of proteins are indicated on the dotted line.

sequence TVTYQPNGKIPVDPNGDPIP. Similarly, LJ1839 contains six repeats of the related sequence TVTYNPNGHIIPVGPDGKPIP in its C-terminal half. The four 80-residue regions of LJ0047 and LJ1839 that begin with these repeats show ≈65% sequence similarity to residues 10–90 of cell-surface bull sperm binding protein SP18 (GenBank accession no. AF129872) or HR44 of *Homo sapiens* (accession no. X91103). A consensus sequence for the repeat common to all these proteins is TVTYxPNGxPxGxPIPxxPxPxYPT. Finding these MUBs in NCC 533 and seven analogous ones in *L. gasseri* implies an important role for them in adherence and/or GIT persistence. Notably, the draft genomes of the milk-fermenting lactobacilli (*L. bulgaricus* and *L. casei*) lack MUB homologs.

Two other predicted cell-surface proteins with extensive sequence repeats are noteworthy. In LJ0391 and LJ1711, >60 sequential exact repeats of the sequence SESLSNSVSM comprise the 600-residue region preceding their predicted C-terminal cell-wall anchor domain. The many short serine-rich repeats in these proteins are similar to that found in the glycosylated Fap1 fimbrial adhesin of *Streptococcus parasanguis* (38) and GspB platelet binding protein of the *Streptococcus gordonii* (GenBank accession no. AY028381). Interestingly, LJ0391 is preceded by six genes whose products show significant similarity and synteny to the accessory *sec* locus required for export of GspB to the cell surface (Fig. 2) (39). LJ1711 is followed by three predicted glycosyl transferases with good similarity to ones also found in *S. gordonii*. It is plausible that LJ1711 and LJ0391 may encode glycosylated cell-surface adhesive or fimbrial proteins analogous to Fap1 (40) and GspB. Preliminary electron micrographs of an exopolysaccharide (EPS)-deficient mutant of NCC 533 show the presence of fimbriae-like structures on the surface of a few bacteria (data not shown). Bioinformatic analysis showed that the ≈30-kb region with LJ0382–LJ0394, including the nine-gene predicted fimbrial operon is absent from all other lactobacilli examined. Although this region is also missing from the *L. gasseri* genome, several flanking gene homologs are present in synteny, suggesting that this is another "unique" insertion into the genomic "backbone" that is shared by *L. gasseri* and *L. johnsonii* (see Fig. 4, which is published as supporting information on the PNAS web site, and below). However, a similar cluster of genes including *secY2*, *secA2*, and other genes in the LJ0384–0394 region as well as a flanking gene encoding a large serine-rich repeat protein are present in the genomes of *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *Streptococcus agalactiae*.

LJ1680 is a predicted 1,234-residue secreted cell-surface protein with 50% amino acid sequence similarity to IgA proteases only found in several pathogenic streptococci where they are thought to function in immune system evasion. Other than these, no homologs of LJ1680 were detected in >200 microbial genomes searched. LJ1680 may function in hydrolysis of extracellular proteins and peptides, but a role in adhesion to the mucosal surface is also possible (41).

As for possible polysaccharidic adhesion factors, we identified an ORF (LJ0913) with 60% identity to an inulin-producing fructosyltransferase (EC 2.4.1.9) of *L. reuteri* (42), 51% identity to a levansucrase (EC 2.4.1.10) of *Streptococcus mutans* (43), and 48% identity to *L. reuteri* levansucrase (AF465251). This suggests that *L. johnsonii* NCC 533 produces fructan. This ability appears to be conserved because PCR analysis of 15 *L. johnsonii* isolates showed the presence of LJ0913 (not shown) and *L. gasseri* contains a homolog with 81% amino acid sequence identity.

The region from LJ1021 to LJ1034 (Fig. 5, which is published as supporting information on the PNAS web site) encodes more than five glycosyltransferases, several sugar nucleotide biosynthetic enzymes, and additional proteins similar to ones encoded in EPS biosynthetic gene clusters of streptococci (44). NCC 533 produces an EPS that is cell-surface attached and also shed into the growth medium, whereas deletion mutants of LJ1021–LJ1034 fail to produce this EPS (data not shown). Whereas LJ1021–LJ1025, LJ1032 encoding UDP-galactopyranose mutase, and LJ1034 encoding a probable repeat unit transporter are well conserved in *L. gasseri*, the remaining EPS genes are not. However, *L. gasseri* has other genes with different, but analogous predicted functions, suggesting that it makes an EPS with an alternative structure.

Uncommon Genes of *L. johnsonii* NCC 533. *L. johnsonii* and *L. gasseri* are described as being very closely related (27). This is supported by comparison of the 16S ribosomal genes (99.6% identical) and the observation that the DNA sequences of many housekeeping genes of these two bacteria also show a high degree of similarity. For example, *rpoBC* (LJ0332 and LJ0333) and the 9-kb ribosomal protein operon (LJ0338 to LJ0355) are >94% and >98% identical in DNA sequence, respectively. However, BLAST analyses (using <30% amino acid identity; 50% alignment cut-off) revealed at least 150 NCC 533 ORFs (exclusive of prophages and IS) that were clearly absent from the draft genome of *L. gasseri* (Fig. 1, outer ring). More than 95% of these ORFs were confirmed as NCC 533-specific by hybridizing whole genome microarrays with labeled *L. gasseri* DNA (data not shown). Many of these genes are clustered and have deducible functions. One example is the cluster LJ0854–LJ0858 (Fig. 3), which contains two different β-galactosidases, a LacI-type regulator, and a lactose/proton symporter. These genes are flanked by UDP-galactose epimerase (GalE), galactokinase (GalK), and galactose-1-phosphate uridyltransferase (GalT). The 13-kb region of the NCC 533 genome centered around *galE* shows exact synteny with the *L. gasseri* genome, except for the clear "insertion" of

LJ0854–LJ0858 into the *L. johnsonii–L. gasseri* "genomic backbone" (Fig. 3). LJ0858 is an uncommon type of β-galactosidase with very few detectable orthologs. It is plausible that LJ0854–LJ0858 confers on NCC 533 the ability to use uncommon β-galactosides encountered in its GIT environment. Also in this region is a common two-subunit β-galactosidase for hydrolysis of lactose. Although *L. gasseri* lacks these β-galactosidases, it does have two phospho-β-galactosidases and two PTS transporters for lactose, which in turn are absent from *L. johnsonii*. However, a small gene remnant of one of the phospho-β-galactosidases is found in a syntenic region of the NCC 533 genome, implying that it may have lost genes for metabolizing lactose via a PTS and phospho-β-galactosidase and acquired genes that encode for direct utilization of lactose.

The LJ0730–LJ0738 region appears to be another *L. johnsonii*-specific gene cluster that confers on NCC 533 the ability to acquire sugars from unusual polysaccharides. This region is absent from the *L. gasseri* genome, although the flanking regions, containing a PTS-transporter system and three β-galactosidases, are present with exact synteny. The predicted functions of the genes in this cluster are as follows: LJ0734, sugar-responsive repressor; LJ0735, a β-glucoside kinase; LJ0736, a phospho-β-glucosidase; LJ0737, a xylQ-type isoprimeverose releasing xylosidase; and LJ0738, another uncommon β-galactosidase similar to LJ0858.

A third NCC 533 metabolic gene cluster not present in *L. gasseri* comprises LJ0635, a maltose-6-phosphate glucosidase, LJ0636, a maltose specific IIBC PTS transporter component, and LJ0637, a RpiR-type phosphosugar-responsive regulator. Another is LJ1257–LJ1268 predicted to encode an uncommon two subunit transketolase (EC 2.2.1.1; LJ1266 and LJ1267), a glycerol kinase, sugar isomerase, and β-glucoside-specific PTS system. A final example of a notable *L. johnsonii*-specific metabolic cluster is LJ1654–LJ1661, containing predicted ribose and glycerol binding proteins, a PTS transporter, and a protein (LJ1661) predicted to be a deoxyribose-phosphate aldolase (EC 4.1.2.4). This enzyme converts 2-deoxyribose-5-phosphate to glyceraldehyde-3-phosphate, suggesting that this cluster may encode utilization of exogenous deoxyriboses. Orthologs of this aldolase are present in only *L. casei*, *Bacteriodes thetaiotamicron*, and *S. mutans*.

Genes Potentially Involved In GIT Interactions. Bile salt hydrolase (BSH), the enzyme releasing taurine or glycine from bile, is almost exclusively associated with GIT-colonizing bacteria, and may impart a selective advantage in the GIT environment (45). This hypothesis is supported by the recent observations that loss of BSH activity by *L. monocytogenes* reduces its survival in the guinea pig GIT by >4 logs, while increasing BSH copy number enhances survival by 10-fold (46). NCC 533 encodes three BSHs, the largest number found in any bacterial genome to date except for *L. plantarum* WCFS1, which has four (47). Other than the previously described cbsHβ (48) and cbsHα (49), the NCC 533 genome contains a third enzyme, LJ1147, with ≈50% identity to BSH of *L. plantarum* (50). Adjacent to one of its BSH encoding genes, NCC 533 has two ORFs (LJ0057/0058) with 80% sequence identity that have been shown to function in uptake of bile (48). Homologs of these transporters were not found in >200 genomes searched, except for the closely related probiotic bacteria *L. gasseri* and *L. acidophilus* (GenBank accession no. AF091248). The multiplicity of BSH encoding genes and bile transporters in NCC 533 implies the potential importance of this gene set for GIT survival and persistence. The mechanism by which such genes enhance survival is unknown but could relate to elevated resistance to bile toxicity or their potential role in incorporation of cholesterol into the cell membrane (51).

The production of bacteriocins by probiotic bacteria could provide a competitive advantage in the intestinal ecosystem and possibly enhance their antipathogenic activities. Some *L. johnsonii* strains have been shown to produce lactacin F, a two-component class II peptide bacteriocin (52). NCC 533 has an almost identical and syntenic lactacin F operon, composed of the bacteriocin structural genes (*lafAX*), an immunity component (*lafI*), and ancillary regulatory and export genes that encode a two-component regulatory system and two ABC exporters (Fig. 6*A*, which is published as supporting information on the PNAS web site). However, IS element IS*Ljo4* is located in the histidine kinase gene (Fig. 6*B*) and may alter or eliminate lactacin F production.

Discussion

Despite their importance in human health, relatively few GIT commensals have had their genomes sequenced. But for those that have, genome analysis has revealed a marked physiological and metabolic specialization that seemingly reflects conditions in their GIT niches. In the case of *L. johnsonii*, the genome sequence shows that this commensal, and its close relative *L. gasseri*, are deficient in biosynthesis of amino acids, purine nucleotides, and cofactors. In apparent compensation, *L. johnsonii* has an impressive array of transporters, peptidases, and proteases. Additionally, the abundance of PTS sugar transporters and β-galactosidases implies heavy reliance on mono-, di-, and trisaccharides to its fuel fermentative metabolism. The predicted physiology of *L. johnsonii* is in striking contrast to genome sequence-based predictions for other GIT commensals such as *L. plantarum* (47), *B. longum* (53), and *B. thetaiotamicron* (54), which make their own amino acids, nucleotides, and many cofactors. Moreover, for energy metabolism *Bifidobacteria* and *Bacteroides* focus more on hemicelluloses and other complex carbohydrates that transit the upper GIT intact and "accumulate" in the distal colon where these two species predominate (55). Because *L. johnsonii* is auxotrophic and lacks enzymes for depolymerization of many complex carbohydrates, effective competition with *Bifidobacteria*, *Bacteroides*, and other bacteria in the colon seems unlikely. Rather *L. johnsonii* appears better adapted to life in the upper GIT, where amino acids, peptides, and lower order oligosaccharides abound. In fact, Marteau *et al.* (7) reported that members of *Enterobacteriaceae* and *Lactobacilliales* are the dominant microbiota in the cecum at the ileal–colonic junction. The metabolic shortcomings of *L. johnsonii* and *L. gasseri* place major restrictions on ecological niches available to them. Indeed, *L. johnsonii* has almost exclusively been detected in human and animal GITs. However, it remains to be seen whether *L. johnsonii* is an "obligate" GIT commensal or has other reservoirs.

Comparison of the *L. johnsonii* NCC 533 and *L. gasseri* ATCC 33323 genomes showed a much closer evolutionary relationship than implied by their reported 35% DNA–DNA homology (8, 27) (Fig. 1, outer ring). Our analyses showed >94% DNA sequence identity between many housekeeping genes of these two species. Additionally, 1,364 orthologous pairs ($E > 10^{-2}$, >70% alignment) were detected in the *L. johnsonii* and *L. gasseri* genomes. In contrast, NCC 533 has only 990 orthologous pairs with *L. plantarum*, 810 with *L. delbrueckii*, and ≈850 with *Lactococcus lactis*, *L. monocytogenes*, and *Bacillus subtilis*. Finally, genome segment alignments of *L. johnsonii* and *L. gasseri* using ERGO (www.integratedgenomics.com) revealed very long sections of colinearity, sparsely punctuated by multigene insertions/deletions. However, *in silico* comparison of the genome of *L. johnsonii* to those of other LAB identified quite a few gene clusters present only in *L. johnsonii*. Many of these could be envisioned as enhancing competitiveness in the upper GIT, and perhaps relate to its interactions with the host or probiotic properties. Most prominent were several metabolic cassettes that apparently encode for uptake, hydrolysis, and further metabolism of saccharides. The organization of these operonic cassettes

[i.e., glycosyl hydrolase(s) + sugar transporter + negative regulator] is reminiscent of the many found in *B. longum* (53), but their products are clearly targeted toward much simpler sugars as substrates.

Another group of important *L. johnsonii*-specific genes are predicted to encode its extensive array of diverse cell-surface proteins. Standing out among these are three mucus-binding proteins, glycosylated fimbriae, and an IgA protease. These and other predicted surface molecules may promote binding to the GIT mucosal surface and could also impact cells of the immune system underlying the intestinal epithelium. It is interesting to note that in searches of >200 microbial genomes, homologs of the predicted fimbrial operons, and the predicted IgA protease LJ1680 of *L. johnsonii* were found only in pathogens, implying that the properties conferred by these genes may be important to interactions of both commensals and pathogens with their hosts.

Another gene set that likely enhances survival in the GIT are the three BSHs and the unique bile transporters of *L. johnsonii*. These have been proposed to play a fundamental role in survival because most GIT lactobacilli posses BSH activity (42) and in one case have been shown to be essential for GIT persistence (46). Circumstantial evidence suggests that BSHs facilitate incorporation of cholesterol or bile into the bacterial membrane (51), changing its fluidity or charge (56) in a way that could affect sensitivity to α-defensins and other host defense molecules (57). This adaptation could strongly select for commensals with BSHs, while disfavoring pathogens or other transients lacking BSHs. The facile genetic tools and functional assays available for *L. johnsonii*, coupled with its genome sequence, position *L. johnsonii* NCC 533 as a model system with exciting opportunities to explore the roles of BSHs, surface proteins, polysaccharides, and many other genes in commensal interactions with its human host.

We thank J. Hermanns of Lion Bioscience for his commitment to this sequencing project, the Joint Genome Institute of the U.S. Department of Energy for making the draft genomes of *L. gasseri* and other LAB publicly available, Rhodia, Inc. for support of bioinformatics in the North Carolina State University group, and J.-R. Neeser for his constant and enthusiastic support.

1. Hooper, L. V. & Gordon, J. I. (2001) *Science* **292,** 1115–1118.
2. Xu, J. & Gordon, J. I. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 10452–10459.
3. Hooper, L. V., Midtvedt, T. & Gordon, J. I. (2002) *Annu. Rev. Nutr.* **22,** 283–307.
4. Simon, G. L. & Gorbach, S. L. (1986) *Dig. Dis. Sci.* **31,** 147S–162S.
5. Fuller, R. (1989) *J. Appl. Bacteriol.* **66,** 365–378.
6. Mercenier, A., Pavan, S. & Pot, B. (2003) *Curr. Pharm. Des.* **9,** 175–191.
7. Marteau, P., Pochart, P., Dore, J., Bera-Maillet, C., Bernalier, A. & Corthier, G. (2001) *Appl. Environ. Microbiol.* **67,** 4939–4942.
8. Johnson, J. L., Phelps, C. F., Cummins, C. S., London, J. & Gasser, F. (1980) *Int. J. Syst. Bacteriol.* **30,** 53–68.
9. Zoetendal, E. G., von Wright, A., Vilpponen-Salmela, T., Ben Amor, K., Akkermans, A. D. & de Vos, W. M. (2002) *Appl. Environ. Microbiol.* **68,** 3401–3407.
10. Bernet-Camard, M. F., Lievin, V., Brassart, D., Neeser, J. R., Servin, A. L. & Hudault, S. (1997) *Appl. Environ. Microbiol.* **63,** 2747–2753.
11. Haller, D., Blum, S., Bode, C., Hammes, W. P. & Schiffrin, E. J. (2000) *Infect. Immun.* **68,** 752–759.
12. Haller, D., Bode, C., Hammes, W. P., Pfeifer, A. M., Schiffrin, E. J. & Blum, S. (2000) *Gut* **47,** 79–87.
13. Ibnou-Zekri, N., Blum, S., Schiffrin, E. J. & von der, W. T. (2003) *Infect. Immun.* **71,** 428–436.
14. Bernet, M. F., Brassart, D., Neeser, J. R. & Servin, A. L. (1994) *Gut* **35,** 483–489.
15. Neeser, J. R., Granato, D., Rouvet, M., Servin, A., Teneberg, S. & Karlsson, K. A. (2000) *Glycobiology* **10,** 1193–1199.
16. Granato, D., Perotti, F., Masserey, I., Rouvet, M., Golliard, M., Servin, A. & Brassart, D. (1999) *Appl. Environ. Microbiol.* **65,** 1071–1077.
17. Ewing, B. & Green, P. (1998) *Genome Res.* **8,** 186–194.
18. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998) *Genome Res.* **8,** 175–185.
19. Schiex, T., Gouzy, J., Moisan, A. & de Oliveira, Y. (2003) *Nucleic Acids Res.* **31,** 3738–3741.
20. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
21. Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A. & Durbin, R. (1998) *Nucleic Acids Res.* **26,** 320–322.
22. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147,** 195–197.
23. Lowe, T. M. & Eddy, S. R. (1997) *Nucleic Acids Res.* **25,** 955–964.
24. Busch, W. & Saier, M. H., Jr. (2002) *Crit. Rev. Biochem. Mol. Biol.* **37,** 287–337.
25. Ventura, M., Canchaya, C., Pridmore, D., Berger, B. & Brussow, H. (2003) *J. Bacteriol.* **185,** 4603–4608.
26. Elli, M., Zink, R., Reniero, R. & Morelli, L. (1999) *Int. Dairy J.* **9,** 507–513.
27. Fujisawa, T., Benno, Y., Yaeshima, T. & Mitsuoka, T. (1992) *Int. J. Syst. Bacteriol.* **42,** 487–491.
28. Rice, P., Longden, I. & Bleasby, A. (2000) *Trends Genet.* **16,** 276–277.
29. Pederson, J. A., Mileski, G. J., Weimer, B. C. & Steele, J. L. (1999) *J. Bacteriol.* **181,** 4592–4597.
30. Siezen, R. J. (1999) *Antonie Leeuwenhoek* **76,** 139–155.
31. Carraway, K. L., Ramsauer, V. P., Haq, B. & Carothers Carraway, C. A. (2003) *BioEssays* **25,** 66–71.
32. Roos, S. & Jonsson, H. (2002) *Microbiology* **148,** 433–442.
33. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997) *Int. J. Neural Syst.* **8,** 581–599.
34. Sanderson, S., Campbell, D. J. & Shastri, N. (1995) *J. Exp. Med.* **182,** 1751–1757.
35. Ganeshkumar, N., Arora, N. & Kolenbrander, P. E. (1993) *J. Bacteriol.* **175,** 572–574.
36. Fischetti, V. A. (2000) in *Gram-Positive Pathogens*, ed. Fischetti, V. A. (Am. Soc. Microbiol., Washington, DC), pp. 11–24.
37. Tuomola, E. M., Ouwehand, A. C. & Salminen, S. J. (2000) *Int. J. Food Microbiol.* **60,** 75–81.
38. Wu, H. & Fives-Taylor, P. M. (1999) *Mol. Microbiol.* **34,** 1070–1081.
39. Bensing, B. A. & Sullam, P. M. (2002) *Mol. Microbiol.* **44,** 1081–1094.
40. Stephenson, A. E., Wu, H., Novak, J., Tomana, M., Mintz, K. & Fives-Taylor, P. (2002) *Mol. Microbiol.* **43,** 147–157.
41. Weiser, J. N., Bae, D., Fasching, C., Scamurra, R. W., Ratner, A. J. & Janoff, E. N. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 4215–4220.
42. van Hijum, S. A., Geel-Schutten, G. H., Rahaoui, H., van der Maarel, M. J. & Dijkhuizen, L. (2002) *Appl. Environ. Microbiol.* **68,** 4390–4398.
43. Song, D. D. & Jacques, N. A. (1999) *Biochem. J.* **341,** 285–291.
44. Stingele, F., Neeser, J. R. & Mollet, B. (1996) *J. Bacteriol.* **178,** 1680–1690.
45. Tanaka, H., Doesburg, K., Iwasaki, T. & Mierau, I. (1999) *J. Dairy Sci.* **82,** 2530–2535.
46. Dussurget, O., Cabanes, D., Dehoux, P., Lecuit, M., Buchrieser, C., Glaser, P. & Cossart, P. (2002) *Mol. Microbiol.* **45,** 1095–1106.
47. Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O. P., Leer, R., Tarchini, R., Peters, S. A., Sandbrink, H. M., Fiers, M. W., *et al.* (2003) *Proc. Natl. Acad. Sci. USA* **100,** 1990–1995.
48. Elkins, C. A. & Savage, D. C. (1998) *J. Bacteriol.* **180,** 4344–4349.
49. Elkins, C. A., Moser, S. A. & Savage, D. C. (2001) *Microbiology* **147,** 3403–3412.
50. Christiaens, H., Leer, R. J., Pouwels, P. H. & Verstraete, W. (1992) *Appl. Environ. Microbiol.* **58,** 3792–3798.
51. Dambekodi, P. C. & Gilliland, S. E. (1998) *J. Dairy Sci.* **81,** 1818–1824.
52. Allison, G. E., Fremaux, C. & Klaenhammer, T. R. (1994) *J. Bacteriol.* **176,** 2235–2241.
53. Schell, M. A., Karmirantzou, M., Snel, B., Vilanova, D., Berger, B., Pessi, G., Zwahlen, M. C., Desiere, F., Bork, P., Delley, M., *et al.* (2002) *Proc. Natl. Acad. Sci. USA* **99,** 14422–14427.
54. Xu, J., Bjursell, M. K., Himrod, J., Deng, S., Carmichael, L. K., Chiang, H. C., Hooper, L. V. & Gordon, J. I. (2003) *Science* **299,** 2074–2076.
55. Vercellotti, J. R., Salyers, A. A., Bullard, W. S. & Wilkins, D. (1977) *Can. J. Biochem.* **55,** 1190–1196.
56. Peschel, A., Jack, R. W., Otto, M., Collins, L. V., Staubitz, P., Nicholson, G., Kalbacher, H., Nieuwenhuizen, W. F., Jung, G., Tarkowski, A., *et al.* (2001) *J. Exp. Med.* **193,** 1067–1076.
57. Wilson, C. L., Ouellette, A. J., Satchell, D. P., Ayabe, T., Lopez-Boado, Y. S., Stratman, J. L., Hultgren, S. J., Matrisian, L. M. & Parks, W. C. (1999) *Science* **286,** 113–117.
58. Altermann, E. & Klaenhammer, T. R. (2003) *OMICS* **7,** 161–169.
59. Pedersen, A. G., Jensen, L. J., Brunak, S., Staerfeldt, H. H. & Ussery, D. W. (2000) *J. Mol. Biol.* **299,** 907–930.