RESEARCH ARTICLE

Genome Sequence and Comparative Genome Analysis of *Lactobacillus casei*: Insights into Their Niche-Associated Evolution

Hui Cai, Rebecca Thompson,† Mateo F. Budinich,* Jeff R. Broadbent,† and James L. Steele**

*Department of Food Science, University of Wisconsin; and †Department of Nutrition and Food Sciences, Utah State University

Lactobacillus casei is remarkably adaptable to diverse habitats and widely used in the food industry. To reveal the genomic features that contribute to its broad ecological adaptability and examine the evolution of the species, the genome sequence of *L. casei* ATCC 334 is analyzed and compared with other sequenced lactobacilli. This analysis reveals that ATCC 334 contains a high number of coding sequences involved in carbohydrate utilization and transcriptional regulation, reflecting its requirement for dealing with diverse environmental conditions. A comparison of the genome sequences of ATCC 334 to *L. casei* BL23 reveals 12 and 19 genomic islands, respectively. For a broader assessment of the genetic variability within *L. casei*, gene content of 21 *L. casei* strains isolated from various habitats (cheeses, $n = 7$; plant materials, $n = 8$; and human sources, $n = 6$) was examined by comparative genome hybridization with an ATCC 334-based microarray. This analysis resulted in identification of 25 hypervariable regions. One of these regions contains an overrepresentation of genes involved in carbohydrate utilization and transcriptional regulation and was thus proposed as a lifestyle adaptation island. Differences in *L. casei* genome inventory reveal both gene gain and gene decay. Gene gain, via acquisition of genomic islands, likely confers a fitness benefit in specific habitats. Gene decay, that is, loss of unnecessary ancestral traits, is observed in the cheese isolates and likely results in enhanced fitness in the dairy niche. This study gives the first picture of the stable versus variable regions in *L. casei* and provides valuable insights into evolution, lifestyle adaptation, and metabolic diversity of *L. casei*.

Introduction

The availability of microbial genomes has allowed for new and remarkable insights into the evolution of microorganisms. Microorganisms evolve via three distinct processes: 1) modification of existing genes via mutation followed by vertical inheritance (Sokurenko et al. 1998; Giraud et al. 2001; Tenaillon et al. 2001; Feldgarden et al. 2003), 2) gain of genes that confer a fitness benefit (Lawrence 1999; De Koning et al. 2000; Ochman et al. 2000; Mclysaght et al. 2003; Springael and Top 2004), and 3) loss of genes that no longer confer a fitness benefit (Cole et al. 2001; Ogata et al. 2001; Mirkin et al. 2003). Modification of existing genes is primarily achieved by random mutations, which occur due to mistakes in DNA replication or mutagenic environmental conditions (Elena and Lenski 2003). This process may be accompanied by gene duplication allowing for retention of parental gene function while a new function evolves (Ohta 2003; Saito et al 2003). Modification of existing genes is also caused by insertion of insertion sequence (IS) elements, which may lead to gene inactivation and chromosomal rearrangements (Top and Springael 2003; Schneider and Lenski 2004), events that are more prevalent in IS-abundant organisms such as *Lactobacillus helveticus* (Callanan et al. 2008) and *Lactobacillus casei*. Those mutations that benefit the organism may be maintained and passed on to succeeding generations, such as the mutations in internalin A (*inlA*), which resulted in increased invasiveness of *Listeria monocytogenes* (Orsi et al. 2007). Bacteria also evolve by large-scale changes in their genome composition. Additions to gene inventory typically involve acquisition of gene clusters from other bacteria through horizontal gene transfer (HGT). Such additions can significantly expand a bacterium's potential for adaptation to a new niche (Lawrence 1999; De Koning et al. 2000; Ochman et al. 2000; Springael and Top 2004). For example, acquisition of an 80-kb pathogenicity island was crucial for emergence of *Vibrio parahaemolyticus* as a human pathogen (Izutsu et al. 2008). Once acquired, these new genes, along with the preexisting genes, undergo selection in the new niche. In this process, genes that no longer confer a fitness benefit, or are detrimental to an organism's new lifestyle, are likely to be lost (Cole et al. 2001; Mirkin et al. 2003). For example, as *Shigella* spp. evolved from *Escherichia coli* to become pathogens, several ancestor traits such as lysine decarboxylase (E.C 4.1.1.18) were lost. Products of lysine decarboxylase greatly inhibit the enterotoxin activity of *Shigella*, so loss of this gene resulted in an organism with enhanced fitness to its new pathogenic lifestyle (Maurelli et al. 1998).

Lactic acid bacteria (LAB) constitute a group of Gram-positive, nonsporing, nonrespiring bacteria that are often involved in food and feed fermentations (Axelsson 2004). They produce lactic acid as their major fermentation end product (Axelsson 2004). The largest and most diverse genus of LAB is *Lactobacillus*, which holds more than 125 species and encompasses a wide variety of organisms, including dairy exclusive organisms (e.g., *Lactobacillus delbrueckii* ssp. *bulgaricus* and *L. helveticus*), organisms commonly found in vertebrate gastrointestinal tracts (e.g., *Lactobacillus acidophilus* and *Lactobacillus gasseri*), and organisms with remarkable adaptability to diverse habitats (e.g., *Lactobacillus plantarum* and *L. casei*) (Kandler and Weiss 1986).

Comparative genomics has facilitated our understanding of LAB evolution. Beginning with the genome sequencing of *L. plantarum* WCFS1 in 2003 (Kleerebezem et al. 2003), over 15 *Lactobacillus* genomes representing more than a dozen species have become publicly available. Comparative analysis of nine LAB genomes by Makarova et al. (2006) indicated that a combination of gene gain and gene

Key words: comparative genome hybridization, evolution, niche adaptation.

E-mail: jlsteele@wisc.edu.

Genome. Biol. Evol. 1:239–257.
doi:10.1093/gbe/evp019
Advance Access publication July 14, 2009

loss occurred during the evolution of these bacteria with various environmental habitats. For several LAB species, these events involved a shift in primary habitat from dynamic and nutritionally variable environments, such as the human gastrointestinal tract and plant materials, to the relatively constant and nutrient-rich dairy niche. Adaptation to the dairy niche has been associated with a trend toward metabolic simplification, resulting in loss of carbohydrate metabolic, amino acid biosynthetic, and cofactor biosynthetic genes as well as an increase in genes for peptide transport and hydrolysis (Bolotin et al. 2004; Hols et al. 2005; van de Guchte et al. 2006; Callanan et al. 2008). For example, *L. delbrueckii* ssp. *bulgaricus* and *L. helveticus*, organisms used in yogurt and cheese manufacture, respectively, have undergone massive gene decay (van de Guchte et al. 2006; Callanan et al. 2008). Approximately 12% and 19% of the *L. delbrueckii* ssp. *bulgaricus* and *L. helveticus* genes, respectively, are pseudogenes. Additionally, relatively few enzymes involved in the amino acid biosynthesis are present in the *L. delbrueckii* ssp. *bulgaricus* and *L. helveticus* genomes, suggesting their adaptation to the protein rich dairy niche (van de Guchte et al. 2006; Callanan et al. 2008). For organisms commonly found in gastrointestinal tract, genetic traits that likely contribute to the organisms' gastric survival and promote interactions with the intestinal mucosa and microbiota have also been identified. For example, in silico analyses of the *L. acidophilus* and *L. gasseri* genomes have predicted the presence of 5 and 14 mucin-binding proteins, respectively, and gene clusters for transport of a diverse group of carbohydrates (Altermann et al. 2005; Azcarate-Peril et al. 2008). The genome of the versatile organism *L. plantarum* contains a large number of regulatory and transport functions, including 25 complete phosphoenolpyruvate sugar–transferase systems (PTS) carbohydrate transport systems (Kleerebezem et al. 2003). Additionally, two lifestyle adaptation regions with unusual GC composition were identified in *L. plantarum*, suggesting their recent acquisition via HGT (Molenaar et al. 2005). Overall, comparative genomics suggests that evolution of LAB has been driven by two major processes: gain of new functions by HGT and loss of dispensable ancestral functions.

Lactobacillus casei is a versatile LAB that has been isolated from a variety of environmental habitats, including raw and fermented dairy (especially cheese) and plant materials (e.g., wine, pickle, silage, and kimchi) as well as the reproductive and gastrointestinal tracts of humans and animals (Kandler and Weiss 1986). The population structure within the *L. casei* species has been analyzed by multilocus sequence typing (MLST) and determined to diverge into three major lineages approximately 1.5 million years ago (Cai et al. 2007). Of particular interest is the very recent divergence (~50,000 years ago) of a cheese cluster, which is consistent with the emergence of this relatively new ecological niche (Cai et al. 2007). Additionally, *L. casei* have extensive and diverse applications in the food industry. They are used as acid-producing starter cultures in milk fermentations, as adjunct cultures for intensification, and as acceleration of flavor development in bacterial-ripened cheeses and are commonly the dominant species of nonstarter lactic acid bacteria in ripening cheese (Mayra-Makinen and Bigret 1998). Also, some *L. casei* strains are utilized as probiotics. The Food and Agriculture Organization (FAO) and the World Health Organization (WHO) of the United Nations defined probiotics as "live microorganisms which when administered in adequate amounts confer a health benefit on the host" (FAO/WHO 2002). *Lactobacillus casei* probiotics have an annual market value of ~6 billion US dollars, with over 10^{20} live cells of *L. casei* consumed annually by humans. Their remarkable adaptability, evolving population structure, and extensive industrial utility have made *L. casei* an organism of significant interest in the scientific community.

The complete genome sequence of *L. casei* ATCC 334, an Emmental cheese isolate, has been reported previously with eight other LAB genomes (Makarova et al. 2006). In this study, we provide a more detailed analysis of this genome and a comparison of the ATCC 334 genome with those of other lactobacilli. Recently, the complete genome sequence of *L. casei* BL23 became available, allowing for initial studies into the core versus variable regions in *L. casei* genome. To gain a broader view of the genetic variability within the *L. casei* species, we have conducted a comparative genome hybridization (CGH) analysis against a collection of 21 strains from a variety of environmental habitats (cheese, $n = 7$; plant material, $n = 8$; and human source, $n = 6$). This study provides valuable insights into the evolution, lifestyle adaptation, and metabolic diversity of *L. casei*.

Materials and Methods

Genome Characterization

A general description of the *L. casei* ATCC 334 genome (GenBank accession number CP000423) was previously reported by Makarova et al. (2006). Gene annotations used in this study were derived using the ERGO package (Integrated Genomics) and National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/). The genome atlas was generated using the implemented program from the Joint Genome Institute (http://img.jgi.doe.gov/cgi-bin/pub/main.cgi). All LAB genome sequences examined were downloaded from the NCBI database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria). Comparative genomic analyses for Clusters of Orthologous Groups (COGs) were carried out using the similarity clustering program implemented in ERGO. *L. casei*-specific coding sequences (CDS) were identified by similarity clustering (1.0×10^{-5}) between CDS from 18 *Lactobacillus* and *Pediococcus* genomes (*L. casei* ATCC 334, *L. acidophilus* NCFM, *Lactobacillus brevis* ATCC 367, *L. delbrueckii* spp. *bulgaricus* ATCC BAA-365, *L. delbrueckii* spp. *bulgaricus* ATCC 11842, *Lactobacillus fermentum* IFO 3956, *L. gasseri* ATCC 33323, *L. gasseri* MV-22, *Lactobacillus jensenii* 1153, *L. helveticus* DPC 4571, *Lactobacillus johnsonii* NCC533, *L. plantarum* WCFS1, *Lactobacillus reuteri* JCM 1112, *L. reuteri* F275, *L. reuteri* 100-23, *Lactobacillus sakei* ssp. *sakei* 23K, *Lactobacillus salivarius* ssp. *salivarius* UCC118, and *Pediococcus pentosaceus* ATCC 25745) in ERGO. Each *L. casei*-specific CDS identified was then BLASTed against the NCBI database (http://blast.ncbi.nlm.nih.gov/Blast). CDS with matches from *Lactobacillus*

rhamnosus, *Lactobacillus paracasei*, and other strains of *L. casei* were deleted. Putative sugar and amino acid metabolic pathways were predicted by KEGG (http://www.genome.jp). All *L. casei* genes were analyzed using the SignalP algorithm (http://www.cbs.dtu.dk/services/SignalP/) for signal peptides. Genes containing a signal peptide were further screened for the presence a lipobox by LipPred (http://www.jenner.ac.uk/LipPred/). Genome comparisons were performed using WebACT (Abbott et al. 2005). Unusual base composition was defined as ±5% from average GC % (46.6%) of *L. casei* genome.

Comparative Genome Hybridization

The genome composition of 21 *L. casei* test strains isolated from cheeses (USA, $n = 3$; Australia, $n = 1$; and Denmark, $n = 3$), plant materials (silage, $n = 3$; pickle, $n = 1$; and wine, $n = 4$), and human sources (blood, $n = 3$ and feces, $n = 3$) was examined by CGH, using ATCC 334 as the reference. CGH were performed against an Affymetrix custom microarray designed to include 2,661 (97%) chromosomal and 17 (85%) plasmid CDSs predicted to occur in *L. casei* ATCC 334 as well as all predicted CDSs in the draft *L. helveticus* CNRZ 32 genome (Smeianov et al. 2007). CDSs that were not included in the microarray design were all transposase-encoding genes.

Genomic DNA was extracted using a MasterPure Gram-Positive DNA Purification Kit (Epicentre). Five micrograms of genomic DNA was fragmented and labeled according to instructions for labeling mRNA for antisense prokaryotic arrays (Affymetrix Inc.). Reactions containing 1 μg labeled DNA were used for hybridization at the Utah State University Center for Integrated Biosystems Affymetrix core facility (Logan, UT). Statistical analysis of microarray data was done by R (www.r-project.org). Array images were reduced to intensity values for each probe and all arrays that met acceptable quality control criteria were used for further analysis. Array preprocessing (background correction, normalization, and summarization) was performed using the Robust Multi-array Average method (Bolstad et al. 2003) using Bioconductor (http://www.bioconductor.org). Cutoff values for presence/absence determination were selected empirically based on hybridization signal intensity from negative (*L. helveticus* CNRZ 32) and positive (from hybridization with *L. casei* ATCC 334 DNA) probe controls. Results were clustered and visualized by TIGR MultiExperiment Viewer (TMEV) 4.0 (Saeed et al. 2003), using Euclidean Distance as a distance metric. The Support Tree method of bootstrapping implemented in TMEV was used to test the reliability of the clustering patterns (1,000 bootstrap resamplings).

Phylogenetic Analysis

PCR amplification and DNA sequencing with *ftsZ*, *metRS*, *mutL*, *pgm*, and *polA* for MLST analysis of the 12 additional strains included in this study was performed as described previously (Cai et al. 2007). Multiple sequence alignments were performed using molecular evolutionary genetic analysis (MEGA) software version 4 (http://www.megasoftware.net). A minimum evolution tree for *L. casei* MLST data was reconstructed by using MEGA based on the numbers of parsimoniously informative sites and the results of a bootstrapping test of strain phylogeny (Kumar et al. 2004). The minimum evolution and Neighbor-Joining trees for RNA polymerase subunits, proteolytic enzymes, and lactate dehydrogenases were reconstructed using MEGA. Bootstrap values on the bifurcating branches were based on 1,000 random bootstrap replicates for the consensus tree. Split decomposition analysis was performed using the SplitsTree program (Huson 1998).

To estimate the divergence time in different clusters of *L. casei*, the minimum evolution phylogeny for the 53 strains based on concatenated sequences of the five MLST loci that could be rooted with homologous genes in the closely related species *P. pentosaceus* (>90% nucleotide sequence identity over a minimum alignment length of 90% of both genes) was used. Calculations were based on the number of single-nucleotide substitutions in each strain and the estimated rate of single-nucleotide substitutions between *E. coli* and *Salmonella enterica* of 4.7×10^{-9} per site per year (Doolittle et al. 1996; Lawrence and Ochman 1998).

Data Deposition

MLST data have been deposited in NCBI GenBank under accession numbers FJ770272–FJ770283 (*ftsZ*), FJ770284–FJ770295 (*metRS*), FJ770296–FJ770307 (*mutL*), FJ770308–FJ770319 (*pgm*), and FJ770320–FJ770331 (*polA*). Comparative genome hybridization data have been deposited in Gene Expression Omnibus under accession number GSE15030.

Results and Discussion

Characterization of ATCC 334 Genome

General Genome Features

A general description of the *L. casei* ATCC 334 genome was previously reported by Makarova et al. (2006). It consists of a circular chromosome of 2,895,264 bp and a plasmid (pLSEI1) of 29,061 bp (fig. 1). The general features of the ATCC 334 genome are presented in supplementary table S1 (Supplementary Material online). The genome of *L. casei* BL23 is 3,079,196 bp and devoid of plasmid DNA. The *L. casei* genomes are the second largest *Lactobacillus* genomes sequenced to date. The total number of CDS (2,771 for ATCC 334 and 3,005 for BL23) is within the predicted range (2,700–3,700) for the ancestor of *Lactobacillus* (Makarova et al. 2006). The number of pseudogenes (82 for ATCC 334) is low compared with other sequenced LAB (range from 17 in *Leuconostoc mesenteroides* to 206 in *Streptococcus thermophilus*). The large genome and low number of pseudogenes suggests that *L. casei* ATCC 334 has not undergone extensive genome decay, as has been observed in the dairy exclusive lactobacilli, such as *L. delbrueckii* ssp. *bulgaricus* (van de Guchte et al. 2006) and *L. helveticus* (Callanan et al. 2008).

pLSEI1

A number of plasmids from LAB have been demonstrated to encode multiple important phenotypic traits

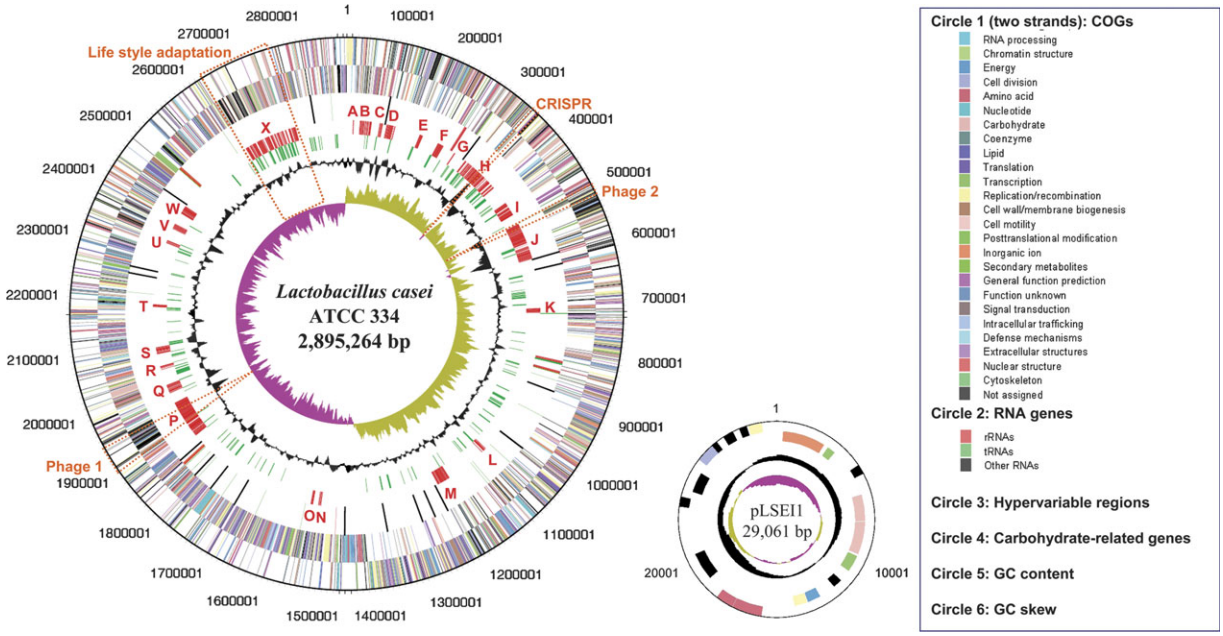


FIG. 1.—Genome atlas of *L. casei* ATCC 334. The color coding of the genomic features in circle 1 represents different COG categories. Locations of CRISPR, phage 1 (Lca1) and phage 2 (phage remnant), the lifestyle adaptation island, and hypervariable regions are labeled.

(McKay 1983; McLandsborough et al. 1995; Yu et al. 1996). Similar to these plasmids, pLSEI1 is predicted to encode multiple phenotypes including lactose hydrolysis (LSEI_A04), transport (LSEI_A05), and regulation (LSEI_A06). LSEI_A04–A06 form an operon-like structure that is flanked by two IS elements, suggesting the possible dissemination of this region by HGT. pLSEI1 also encodes genes for copper export (LSEI_A01) and regulation (LSEI_A02), which are possibly involved in controlling excess accumulation of cytoplasmic copper. Copper resistance may have been important to survival of *L. casei* ATCC 334 in Emmental cheese, from which it was originally recovered, because manufacture of Swiss-type cheeses in copper vats results in high levels of soluble copper (Barre et al. 2007). Additionally, pLSEI1 encodes genes for glutamine transport (LSEI_A10-A12) and an oxidoreductase (LSEI_A08). Given the CDS present in pLSEI1, it is likely that the selection in cheese-related environments has resulted in acquisition and retention of pLSEI1 in ATCC 334.

Prophage, Clustered Regularly Interspaced Short Palindromic Repeats, and Bacteriocin

Two bacteriophage regions are present in ATCC 334 (fig. 1). One appears to be an intact prophage, designated Lca1, which has been previously characterized (Ventura et al. 2006). The other is a ~9.6-kb phage remnant located at 511,120 bp. It contains CDS involved in phage replication and regulation. Overall, prophage-related genes encompass ~1.9% of the ATCC 334 genome, suggesting an important role for lysogenic bacteriophage in *L. casei* evolution.

Clustered regularly interspaced short palindromic repeats (CRISPR) represent a family of DNA repeats typically composed of short and highly conserved repeats, interspaced by variable sequences called spacers, which are often found adjacent to *cas* (CRISPR-associated) genes (Haft et al. 2005; Sorek et al. 2008). CRISPRs have been shown to provide a nucleic acid-based "immunity" against bacteriophage infection, possibly through a RNA interference–like mechanism (Barrangou et al. 2007; Horvath et al. 2008). Analysis of CRISPR loci may provide insights into the coevolution of bacteriophage and their hosts. *Lactobacillus casei* ATCC 334 and BL23 contain two distinct types of CRISPR loci. The CRISPR locus in ATCC 334, designated Lca2, belongs to the diverse Ldbu1 family (Horvath et al. 2008). It contains 22 perfect repeats of a 29-bp long sequence {5′GTTTTCCCCGCACATGCGGGGGTGATCCY(C or T)}. Immediate upstream of the DNA repeats are eight *cas* genes with an average GC content of 57.6% (vs. 46.6% for the chromosome), suggesting their recent acquisition. The CRISPR locus in BL23, designated Lca1, belongs to the well-conserved Lsal1 family. It contains 22 perfect repeats of a 36-bp long sequence (5′GTCTCAGGTAGATGTCGAATCAATCAGTTCAAGAGC). Four *cas* genes are found upstream and two are found downstream of the DNA repeats with an average GC of 43.3% (Horvath et al. 2008), suggesting that either these genes were acquired from an organism with similar GC content or were acquired in the distant past. For both Lca1 and Lca2 CRISPRs, IS elements are found on one side of the CRISPR-*cas* region, suggesting their possible dissemination via HGT.

Bacteriocins, small antimicrobial peptides widely produced by LAB, likely provide their producers with a competitive advantage (Riley and Wertz 2002; De Vuyst and Leroy 2007). The ATCC 334 genome encodes a typical class II bacteriocin gene cluster (Oppegard et al. 2007). This region contains two structural genes (LSEI_2374 and 2375) predicted to encode the preforms of the two peptides that constitute the bacteriocin, which have highest similarity

with those from *L. rhamnosus* (amino acid identity 48%, *e* value 7×10^{-18}), an immunity gene (LSEI_2376) predicted to protect ATCC 334 from this bacteriocin, an ATP-binding cassette (ABC) transporter (LSEI_2384) predicted to transfer the bacteriocin across the membrane and remove the leader sequence, and a gene (LSEI_2381) predicted to encode an accessory protein which also appears to be required for bacteriocin secretion. The presence of the adjacent IS elements suggests the possible acquisition of this region via HGT.

Amino Acid Biosynthesis and Proteolytic Enzyme System

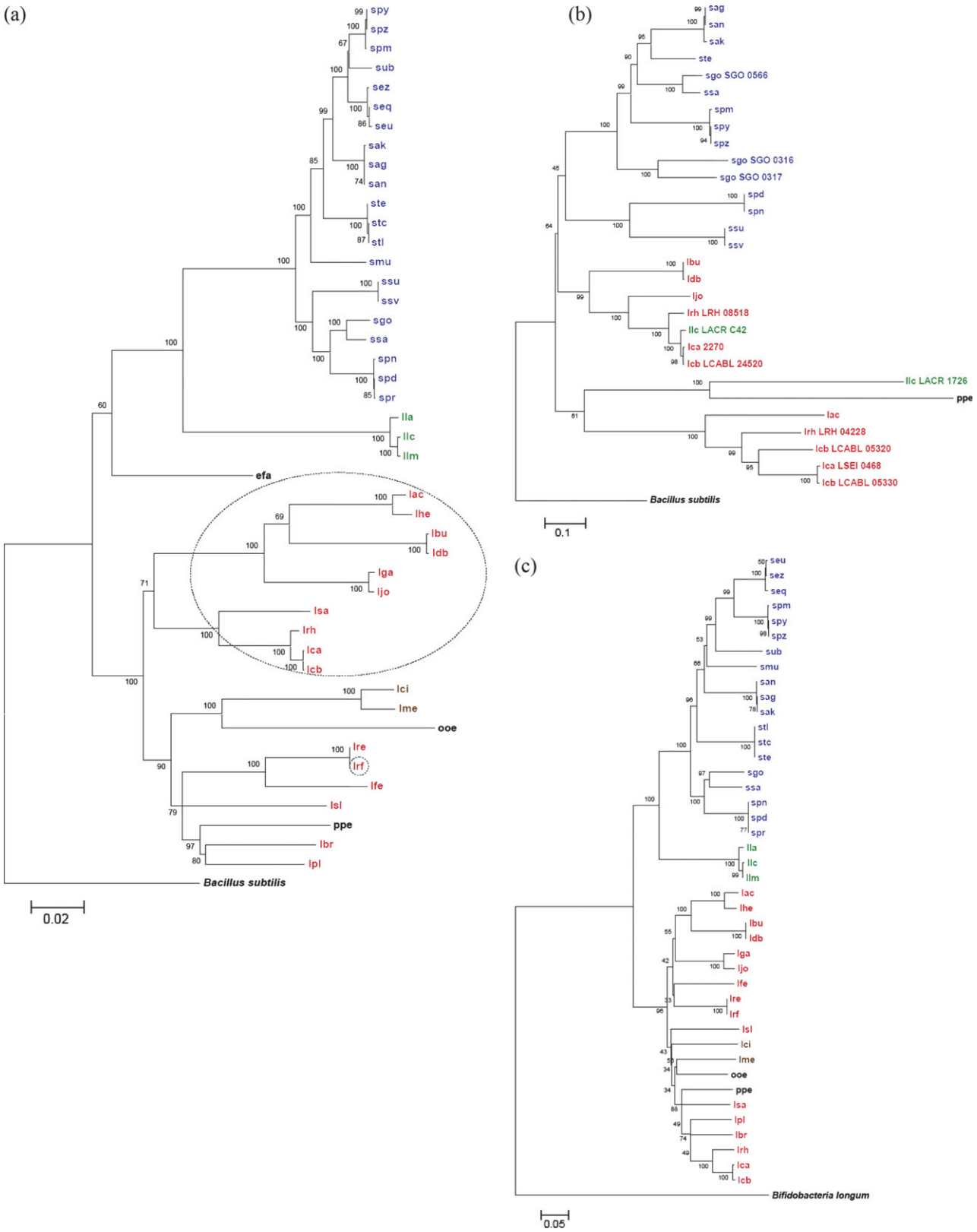
Amino acid biosynthetic pathways are deficient to varying degrees in LAB. For example, dairy organisms *L. delbrueckii* ssp. *bulgaricus* and *L. helveticus* have lost the majority of their amino acid biosynthetic capacities (van de Guchte et al. 2006; Callanan et al. 2008; Christiansen et al. 2008); *L. acidophilus* and *L. gasseri*, organisms commonly found in the human gastrointestinal tract, are auxotrophic for 14 and 17 amino acids, respectively (Altermann et al. 2005; Azcarate-Peril et al. 2008); *L. plantarum*, a versatile organism, is predicted to synthesize all amino acids except for leucine, isoleucine, and valine (Kleerebezem et al. 2003). *L. casei* ATCC 334, like *L. plantarum*, possesses enzymes for biosynthesis of all amino acids except for the branched-chain amino acids (valine, leucine, and isoleucine); this suggests that even though ATCC 334 was isolated from cheese, this strain may be capable of inhabiting a variety of ecological niches, including protein-limited environments.

Although *L. casei* ATCC 334 is capable of synthesizing all but the branched-chain amino acids, it is equipped with a proteolytic enzyme system that allows it to acquire amino acids from proteins present in its environment. This proteolytic enzyme system is likely vital for the acquisition of essential branched-chain amino acids and likely provides the bacterium with a selective advantage in protein-rich environments as it is more energetically favorable to obtain amino acids from environmental proteins than de novo synthesis. Detailed analysis of the proteolytic enzyme system in dairy LAB, most notably *Lactococcus lactis* and *L. helveticus*, has shown that enzymes involved in procurement of amino acids from caseins can be loosely divided into three major categories: 1) an extracellular cell surface–associated proteinase, termed lactocepin (EC 3.4.21.96), that hydrolyzes caseins into oligopeptides; 2) specialized transport systems to take up those oligopeptides, as well as di- and tripeptides, and free amino acids that may be present in the medium; and 3) intracellular endopeptidases and exopeptidases, including many that are specific for proline-containing peptides, which degrade internalized peptides into oligopeptides and, eventually, free amino acids (Christensen et al. 1999; Savijoki et al. 2006).

Lactobacillus casei ATCC 334 grows well in milk and its genome contains three CDS for lactocepin (LSEI_0465, 0468, and 2270), which belong to two different types of lactocepin: type PrtP (LSEI_2270), which was first observed in *L. lactis*, and type PrtR (LSEI_0465 and 0468), first observed in *L. rhamnosus*. However, LSEI_0465 is only the C-terminal fragment of the lactocepin gene. The catalytic domains of this lactocepin gene (PFAM00082 and PFAM06280) are encoded by its upstream pseudogenes LSEI_0466 and LSEI_0467, both of which contain a putative start codon. Additionally, LSEI_0467 is predicted to possess a suitable ribosome-binding site and signal peptide sequence typical of lactocepin. Consequently, LSEI_0465 may not encode a functional extracellular protease. Lactocepins are synthesized as an inactive precursor and some require a membrane-bound lipoprotein, PrtM, a proteinase maturation protein, for autocatalytic maturation (Haandrikman et al. 1991). Lactococcal lactocepin is known to require PrtM (Haandrikman et al. 1991), and LSEI_2270 is an ortholog to the lactococcal PrtP enzyme. The presence of a PrtM ortholog (LSEI_2271) immediately adjacent to LSEI_2270 in ATCC 334 indicates that this enzyme is required for maturation of at least one *L. casei* lactocepin.

Peptides and amino acids that are liberated by the action of proteolytic enzymes may be translocated to the cytoplasm by a variety of peptide and amino acid transporters. Analysis of the ATCC 334 genome revealed the presence of at least 10 ABC-type transporters, capable of translocating amino acids and oligopeptides as well as several di- and tripeptide transport systems. Additionally, ATCC 334 contains a complete oligopeptide transport system. Once internalized, the peptides are degraded by a variety of peptidases. ATCC 334 possesses 27 CDS encoding peptidases (supplementary table S2, Supplementary Material online), which is more than the 19 present in *L. plantarum* (Kleerebezem et al. 2003), the 20 present in *L. acidophilus* (Altermann et al. 2005), and the 24 present in *L. helveticus* (Callanan et al. 2008), but similar to the 26 present in *L. delbrueckii* ssp. *bulgaricus* (van de Guchte et al. 2006) and the 29 present in *L. gasseri* (Azcarate-Peril et al. 2008). Therefore, the number of peptidases does not differ significantly among lactobacilli.

To probe the evolution of the proteolytic enzyme system of *L. casei*, phylogenetic trees for lactocepin and four peptidases (aminopeptidase N, PepN; endopeptidase O, PepO; X-prolyl dipeptidyl aminopeptidase, PepX; and aminopeptidase C or endopeptidase E, PepC/E) from 45 completely sequenced LAB strains, which represent 29 LAB species, were constructed (fig. 2) and compared with a reference tree based upon the concatenated alignment of four subunits (α, β, β', and δ) of the RNA polymerase (fig. 2*a*). Single genes for PepN (fig. 2*c*) and PepX (fig. 2*d*) were found in all 45 LAB strains and they clustered in groups related to the reference tree (fig. 2*a*), suggesting that mutation followed by vertical inheritance likely described their evolutionary path. In contrast, multiple copies of *pepC*/*E* (fig. 2*e*) and *pepO* (fig. 2*f*) were identified. It has been demonstrated previously that deletion of 1–4 amino acid residues from the C-terminal end of lactococcal PepC abolished its aminopeptidase activity, resulting in endopeptidase PepE activity (Mata et al. 1997; Mata et al. 1999). The *pepC* and *pepE* genes in figure 2*e* formed two distinct groups. The *pepC*s were present in all LAB examined and clustered in groups related to the reference tree. The highest abundance in *pepE*s was observed in *L. johnsonii*, where two *pepE*s from *L. johnsonii* (LJ0716 GI42518638 and LJ0719) seemed to be evolutionarily distinct from other *pepC*/*E*s examined. Additionally, all representatives of the *L. casei–L.*


(a)
spy
spz
spm
sub
sez
seq
seu
sak
sag
san
ste
stc
stl
smu
ssu
ssv
sgo
ssa
spn
spd
spr
lla
llc
llm
efa
lac
lhe
lbu
ldb
lga
ljo
lsa
lrh
lca
lcb
lci
lme
ooe
lre
lrf
lfe
lsl
ppe
lbr
lpl
Bacillus subtilis
0.02
(b)
sag
san
sak
ste
sgo SGO 0566
ssa
spm
spy
spz
sgo SGO 0316
sgo SGO 0317
spd
spn
ssu
ssv
lbu
ldb
ljo
lrh LRH 08518
llc LACR C42
lca 2270
lcb LCABL 24520
llc LACR 1726
ppe
lac
lrh LRH 04228
lcb LCABL 05320
lca LSEI 0468
lcb LCABL 05330
Bacillus subtilis
0.1
(c)
seu
sez
seq
spm
spy
spz
sub
smu
san
sag
sak
stl
stc
ste
sgo
ssa
spn
spd
spr
lla
llc
llm
lac
lhe
lbu
ldb
lga
ljo
lfe
lre
lrf
lsl
lci
lme
ooe
ppe
lsa
lpl
lbr
lrh
lca
lcb
Bifidobacteria longum
0.05

delbrueckii group of *Lactobacillaceae* (fig. 2*a*) contained *pepE*, and *pepE* was present in only one LAB genome (*L. reuteri* JCM1112) outside of this group. These results suggested that *pepC* duplication and *pepC*/*pepE* conversion may have occurred before the divergence of the *L. casei–L. delbrueckii* group of *Lactobacillaceae* from *Leuconostocaceae* and the *Pediococcus–L. plantarum* group of *Lactobacillaceae*. As to *pepO*s, characterization of the three *pepO*s from *Lb. helveticus* CNRZ32 has demonstrated that *pepO* differs from *pepO2* and *pepO3* in its substrate selectivity (Chen et al. 2003; Sridhar et al. 2005). Interestingly, *pepO* duplication was also observed only in the *L. casei–L. delbrueckii* group of *Lactobacillaceae*. Overall, *pepC*/*E* and *pepO* may have evolved by gene duplication followed by mutation and vertical inheritance to provide for distinct new functions. Such evolution may have expanded the endopeptidase complements for the *L. casei–L. delbrueckii* group of *Lactobacillaceae*, thereby providing these organisms with a selective advantage in protein-rich environments.

Comparison of the RNA polymerase-based LAB reference tree (fig. 2*a*) with the lactocepin tree (fig. 2*b*) revealed that one of the *L. lactis* subsp. *cremoris* SK11 lactocepin genes (LACR_C42) did not cluster in groups related to the reference tree. Instead, LACR_C42 clustered with lactocepins from *L. casei* ATCC 334 (LSEI 2270) and *L. casei* BL23 (LCABL 24520). Unlike the other lactococci lactocepin examined (LACR_1726), LACR_C42 is plasmid encoded. This aberrant clustering is likely the result of recombination between LACR_C42 and the lactocepins from *L. casei*, as indicated by an interconnected network structure identified by Split decomposition analysis (supplementary fig. S1, Supplementary Material online). This result suggests that plasmid-mediated HGT followed by recombination may have contributed to the evolution of lactocepins in LAB.

Carbohydrate Transport and Metabolism

Lactobacillus casei ATCC 334 contains a large number of CDS involved in carbohydrate utilization. Previously, we determined that ATCC 334 can utilize ribose, galactose, glucose, fructose, sucrose, mannitol, mannose, *N*-acetyl glucosamine, tagatose, cellobiose, maltose, lactose, trehalose, turanose, salicin, melezitose, and inulin (Cai et al. 2007). The PTS systems in *L. casei* have been reported as previously described (Monedero et al. 2007). In silico analysis of the ATCC 334 genome revealed 21 complete PTS and several incomplete PTS systems. Putative PTS genes are identified for mono-, di-, and polysaccharides, including galactose, galactitol, glucose, fructose, mannitol, sorbitol, mannose, N-acetyl galactosamine, cellobiose, lactose, sucrose, trehalose, and arbutin. Similar to *L. plantarum* (Kleerebezem et al. 2003), most carbohydrate-related CDS in ATCC 334 are clustered near the origin of replication (fig. 1). The ability to utilize such a variety of carbohydrates, including many that are not present in milk-based environments, supports the hypothesis that ATCC 334 is capable of inhabiting a variety of ecological niches.

Lactobacillus casei produces lactic acid from hexose sugars via the Embden–Meyerhof Pathway and phosphoketolase pathway, leading to homolactic and heterolactic fermentation profiles, respectively (Kandler 1983). Additionally, pentoses can be metabolized via transketolase (pentose phosphate cycle) and phosphoketolase pathways, leading to homolactic and heterolactic fermentation profiles (Tanaka et al. 2002). The CDS for these intact pathways were found in ATCC 334. Regarding pyruvate metabolism, both D- and L-lactate dehydrogenase (*ldhD* and *ldhL*) CDS are present, which convert pyruvate into D- and L-lactate, respectively. A remarkable degree of redundancy is observed for these CDS, including at least six *ldh* CDS in ATCC 334 and five *ldh* CDS in BL23 (Rico et al. 2008); however, substrate selectivity of these enzymes remains unknown. Neighbor-Joining trees for *ldhD* and *ldhL* from 15 *Lactobacillus* strains, representing 13 *Lactobacillus* species, were constructed (supplementary fig. S2, Supplementary Material online). In general, *L. casei ldh* genes grouped together with those from *L. rhamnosus*, suggesting these genes are evolutionarily closely related, and the redundancy of these genes is achieved via gene duplication followed by vertical inheritance. Dehydrogenases are essential for the regeneration of NAD+, hence they have a key role in maintaining cellular redox balance and metabolic flux (Hoefnagel and Starrenburg 2002; Garvie 1980). Additionally, there is a large number of other pyruvate dissipating enzymes predicted to catalyze the production of a variety of

←

FIG. 2.—Minimum evolution phylogenetic trees for (*a*) concatenated alignment of four subunits (α, β, β′, and δ) of the RNA polymerase, (*b*) lactocepin, (*c*) PepN, (*d*) PepX, (e) PepC/E, and (*f*) PepO from 45 LAB strains. *Bacillus* and *Bifidobacteria* are used as outgroups. Bootstrap values on the bifurcating branches are based on 1,000 random bootstrap replicates for the consensus tree. Strains that contain both PepC and PepE are circled in the RNA polymerase tree (*a*). Strains representing different genera are color coded; *Lactobacillus* is shown in red, *Streptococcus* in blue, *Lactococcus* in green, *Leuconostoc* in brown, and the rest (*Oenococcus*, *Enterococcus*, and *Pediococcus*) in black. Gene ID is given for strains with more than one homolog. Strain code: efa, *Enterococcus faecalis* V583; lac, *L. acidophilus* NCFM; lbr, *L. brevis* ATCC 367; lca, *L. casei* ATCC 334; lcb, *L. casei* BL23; ldb, *L. delbrueckii* ATCC 11842; lbu, *L. delbrueckii* ATCC BAA-365; lfe, *L. fermentum* IFO3956; lga, *L. gasseri* ATCC 33323; lhe, *L. helveticus* DPC 4571; ljo, *L. johnsonii* NCC 533; lpl, *L. plantarum* WCFS1; lre, *L. reuteri* DSM 20016; lrf, *L. reuteri* JCM 1112; lrh, *L. rhamnosus* HN001; lsa, *L. sakei* subsp. *sakei* 23K; lsl, *L. salivarius* UCC118; llm, *L. lactis* subsp. *cremoris* MG1363; llc, *L. lactis* subsp. *cremoris* SK11; lla, *L. lactis* subsp. *lactis* IL1403; lci, *Leuconostoc citreum* KM20; lme, *L. mesenteroides* subsp. *mesenteroides* ATCC 8293; ooe, *Oenococcus oeni* PSU-1; ppe, *P. pentosaceus* ATCC 25745; sag, *Streptococcus agalactiae 2603* (serotype V); sak, *S. agalactiae* A909 (serotype Ia); san, *S. agalactiae* NEM316 (serotype III); seu, *Streptococcus equi* subsp. *equi* 4047; seq, *Streptococcus equi* subsp. *zooepidemicus*; sez, *S. equi* subsp. *zooepidemicus* MGCS10565; sgo, *Streptococcus gordonii* str. Challis substr. CH1; smu, *Streptococcus mutans* UA159; spd, *Streptococcus pneumoniae* D39; spr, *S. pneumoniae* R6; spn, *S. pneumoniae* TIGR4; spz, *Streptococcus pyogenes* MGAS5005 (serotype M1); spm, *S. pyogenes* MGAS8232 (serotype M18); spy, *S. pyogenes* SF370 (serotype M1); ssa, *Streptococcus sanguinis* SK36; ssu, *Streptococcus suis* 05ZYH33; ssv, *S. suis* 98HAH33; stc, *S. thermophilus* CNRZ1066; ste, *S. thermophilus* LMD-9; stl, *S. thermophilus* LMG18311; and sub, *Streptococcus uberis* 0140J.

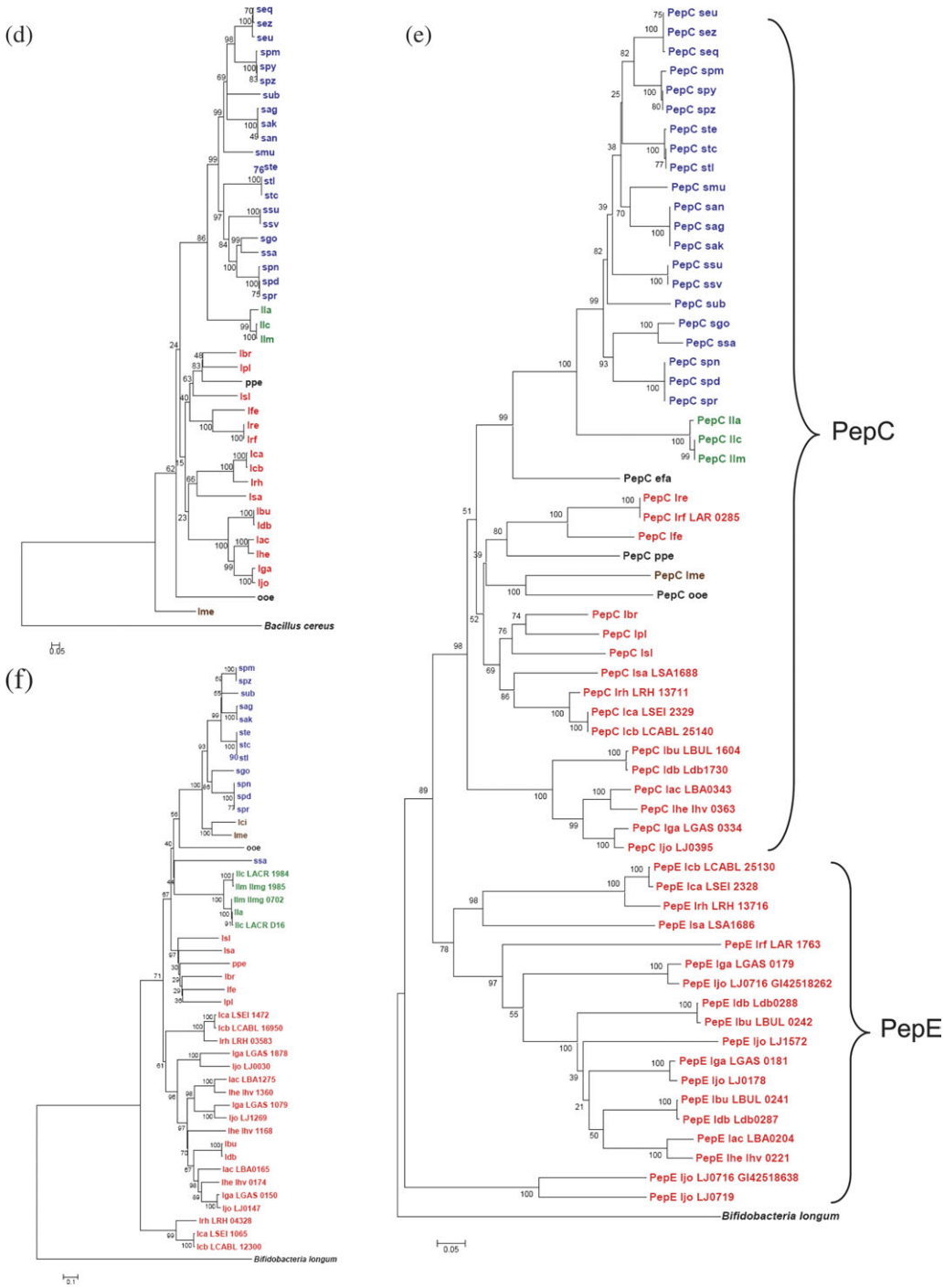


FIG. 2.—continued.

metabolites including formate, acetate, ethanol, acetoin, and 2,3-butanediol. Recently, Viana et al. (2005) characterized the *ldhL* genes from *L. casei* BL23. Their results demonstrated that these enzyme differ in their substrate selectivity and hence allow the organism to regenerate NAD+ from a wide variety of substrates, thereby providing for greater metabolic diversity and the ability to inhabit an expanded set of ecological niches.

Regulation

ATCC 334 contains 16 complete two-component regulatory systems, the most observed among sequenced lactobacilli. This is significantly higher than the 5 in *L. delbrueckii* ssp. *bulgaricus* and *L. gasseri* and the 4 in *L. helveticus*, higher than the 9 in *L. acidophilus* and similar to the 13 in *L. plantarum*. Furthermore, a total of 124

transcriptional regulators are found in ATCC 334, comprising ~4.5% of the genome. Although lower than what is present in *L. plantarum* (234; 7.7% of total CDS) and similar to what in present in *L. acidophilus* (96; 5.2%) and *L. gasseri* (70; 4.0%), the number of transcriptional regulators in *L. casei* is much higher than that observed in *L. delbrueckii* ssp. *bulgaricus* (53; 3.1%) or *L. helveticus* (44; 2.7%). A possible explanation is that *L. delbrueckii* ssp. *bulgaricus* and *L. helveticus* have adapted to the nutrient-rich milk environments, where less adaptive regulation is required than that required by organisms commonly found in gastrointestinal tract like *L. acidophilus* and *L. gasseri*. The significantly high numbers of transcriptional regulators observed in versatile microorganisms like *L. casei* and *L. plantarum* may reflect their requirements for dealing with diverse environmental conditions. These results indicate that ATCC 334 is capable of sensing and responding to a relatively wide variety of environmental conditions.

Cell Surface and Secreted Components

Bacteria exhibit different cell surface structures according to the ecological niche they inhabit. In addition, probiotic bacteria should be able to adhere to gut epithelial tissue and colonize, at least transiently, the mucosal surfaces in gastrointestinal tract (Mack et al. 2003; Marco et al. 2006). ATCC 334 is predicted to secrete 361 proteins, of which 48 contain lipoprotein signal peptide and are predicted to attach to the cytoplasmic membrane. The *L. casei* ATCC 334 proteins predicted to be involved in the adherence to eukaryotic extracellular macromolecules include two collagen adhesion proteins (PFAM PF05737, LSEI_2511 and 2512), a fibronectin-binding protein (PFAM PF05833, LSEI_1439), and a putative mucin-binding protein (PFAM PF06458, LSEI_2320). Homologs of these CDS have been identified in *L. casei* BL23. There is a trend toward organisms isolated from the gastrointestinal tract containing a greater number of mucin-binding proteins. For example, two gastrointestinal isolates, *L. acidophilus* and *L. gasseri*, contain 5 and 14 copies of mucin-binding protein, respectively; whereas the dairy organisms, *L. delbrueckii* ssp. *bulgaricus*, *L. helveticus*, and the versatile organism, *L. plantarum*, contain 0, 0, and 2, respectively. Although the mucin-binding proteins from *L. acidophilus* and *L. gasseri* contain 6–12 copies of mucin-binding domain (PFAM PF06458), only one copy is found in LSEI_2320. The presence of a single domain-containing putative mucin-binding protein in ATCC 334 suggests that this strain is poorly adapted to the gastrointestinal environmental niche. However, adhesion of *Lactobacillus* strains to Caco-2 epithelial cells is known to be a strain-dependent trait (Tuomola and Salminen 1998). Therefore, whereas ATCC 334 does not appear to be well adapted to the vertebrate gastrointestinal tract, other strains of *L. casei* are likely better adapted to this niche.

Many LAB produce extracellular polysaccharides (EPS) that are either excreted as slime (ropy form) or remain attached to the bacterial cell wall forming capsular EPS (De Vuyst and Degeest 1999). These polymers may be composed of one type of sugar monomer (homopolysaccharide) or consist of several types of monomers (hetero-EPS). Hetero-EPS produced by LAB have been found to influence the functional properties of fermented foods as well as the adhesion properties of probiotic strains (De Vuyst and Degeest 1999; Ruas-Madiedo et al. 2006). Gene clusters for hetero-EPS production have been found in most lactobacilli sequenced to date and typically include CDS encoding regulation, chain-length determination, assembly of the basic repeat unit, polymerization, and translocation (De Vuyst and Degeest 1999). The *L. casei* ATCC 334 genome contains one relatively small cluster that includes genes for the conserved EPS translocase or "flippase" Wzx (LSEI_0238) and five glycosyltransferases (LSEI_0232, 0233, 0234, 0235, and 0239), interrupted by two genes of unrelated function. A similar cluster is present in the BL23 genome, but the latter strain also has a second EPS-related cluster that includes CDS for additional glycosyltransferases, the conserved polysaccharide biosynthesis chain length regulator Wzd and tyrosine-protein kinase Wze, and for the glycosyl-1-phosphate transferase that catalyzes the first step in assembly of the EPS basic repeating unit. This observation suggests that ATCC 334 probably does not produce hetero-EPS and that the gene cluster found in this strain serves an alternative function.

Lactobacillus Interspecies Comparison

Comparative genome analysis of related *Lactobacillus* species allows for the identification of niche- or lifestyle-specific genome characteristics. Comparative genome analysis between *L. casei* ATCC 334 and other sequenced lactobacilli revealed that ATCC 334 contains a relatively high number of IS elements and carbohydrate-related genes. The *L. casei* ATCC 334 genome contains 120 (~4.3% of total CDS) IS elements, which is second in number only to *L. helveticus* (213, ~13.2% of total CDS) among sequenced lactobacilli (Callanan et al. 2008). These IS elements are comprised of 11 different families, nine of which (IS*3*, IS*3* like, IS*4*, IS*5*, IS*6*, IS*30*, IS*110*, IS*256*, and IS*1380*) appear in multiple copies. Among them, IS*3* and IS*30* are among the most numerous elements in the genome, with 26 and 58 copies identified, respectively, and are also highly conserved, indicating recent integration and multiplication events. Overall, the abundance and diversity of IS in ATCC 334 suggests that these elements have an important role in genome evolution of *L. casei*.

A comparison of the five most abundant COG categories, which contain genes that typically have the same functional category in different organisms, of 12 *Lactobacillus* spp. is presented in supplementary figure S3 (Supplementary Material online). The most common COGs for all the organisms examined are those associated with general housekeeping functions (general functions and replication) and unknown functions. Interestingly, the next most abundant COG for *L. casei* was carbohydrate utilization, a feature that is also seen with other versatile (e.g., *L. plantarum*) and gastrointestinal (e.g., *L. acidophilus*, *L. gasseri*, and *L. johnsonii*) LAB species. This result suggests that the ability to utilize a wide variety of carbohydrates is an important attribute for organisms that inhabit the gastrointestinal tract or are able to inhabit a variety of ecological niches.

To identify genes that are unique to *L. casei*, similarity clustering (1.0×10^{-5}) of ATCC 334 CDS was performed

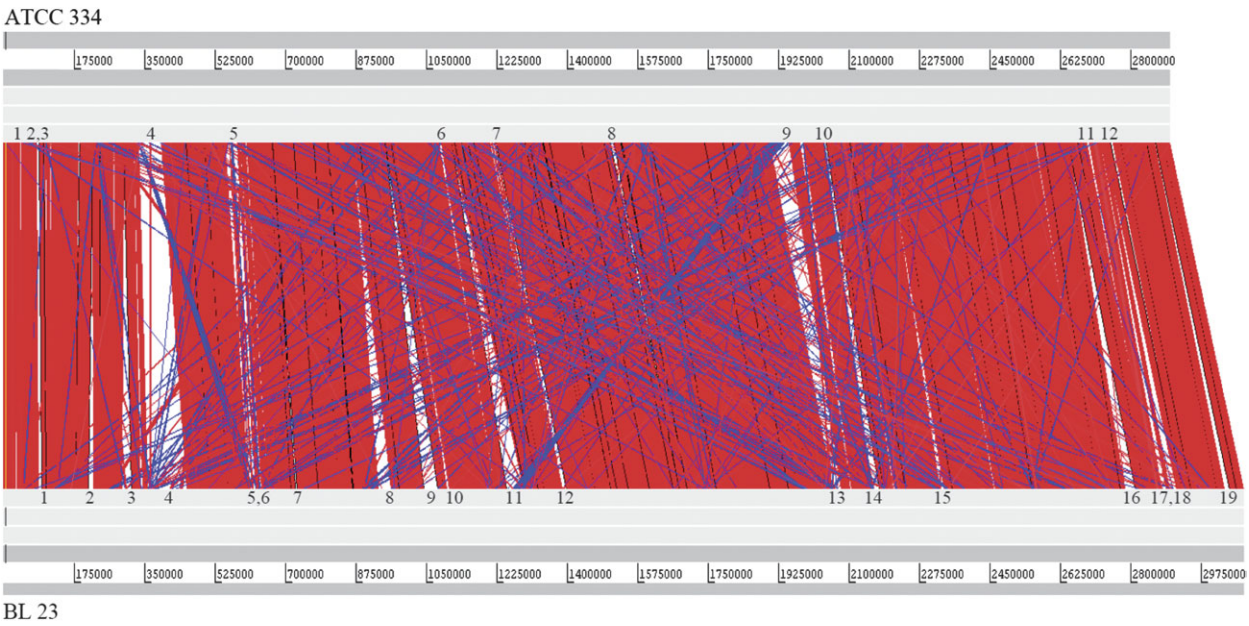


FIG. 3.—Genome comparison between *L. casei* strains ATCC 334 (top) and BL23 (bottom). GIs of more than 5 kb in each genome are numbered. Homologous genomic sequences (BlastN matches) are indicated by red (same orientation) and blue (inversion) lines between the chromosomes.

against 17 sequenced *Lactobacillus* and *Pediococcus* strains (*L. acidophilus* NCFM, *L. brevis* ATCC 367, *L. delbrueckii* spp. *bulgaricus* ATCC BAA-365, *L. delbrueckii* spp. *bulgaricus* ATCC 11842, *L. fermentum* IFO 3956, *L. gasseri* ATCC 33323, *L. gasseri* MV-22, *L. jensenii* 1153, *L. helveticus* DPC 4571, *L. johnsonii* NCC533, *L. plantarum* WCFS1, *L. reuteri* JCM 1112, *L. reuteri* F275, *L. reuteri* 100-23, *L. sakei* ssp. *sakei* 23K, *L. salivarius* ssp. *salivarius* UCC118, and *P. pentosaceus* ATCC 25745). Of the 2,771 predicted *L. casei* ATCC 334 CDS, 197 (7.1%) CDS are non-*Lactobacillus*/*Pediococcus* but *L. casei* specific. Among these 197 CDS (supplementary table S3, Supplementary Material online), 177 encode for hypothetical proteins and 20 encode for metabolic enzymes (e.g., butyrate kinase EC 2.7.2.7; phosphate butyryltransferase EC 2.3.1.19; 6-hydroxy- D-nicotine oxidase EC 1.5.3.6), ABC transporter permeases, and transcriptional regulators, etc. All 197 CDS were screened for homologs using Blast tools in the NCBI database. Eighty-nine CDS do not have a close match in the database. The most common matches (with *e* value equal to or lower than 1.0^{-5}) came from *Enterococcus* (21), *Clostridium* (16), *Streptococcus* (10), and *Bacillus* (10). The high number of matches of *L. casei* CDS with those from *Enterococcus* is even more striking considering that only one complete *Enterococcus* genome is available in the NCBI database, compared with 23, 39, and 25 *Clostridium*, *Streptococcus*, and *Bacillus* genomes, respectively. Like *Lactobacillus*, *Enterococcus* are low GC Gram-positive LAB and are natural inhabitants of the mammalian gastrointestinal tract (Devriese et al. 1992). However, *Enterococcus* are highly promiscuous microbes that contain numerous mobile genetic elements encoding traits such as antibiotic resistance and virulence, which facilitate HGT between related or evolutionary distant microflora (Paulsen et al. 2003). The high number of *L. casei* CDS with matches to *Enterococcus* suggests that HGT from *Enterococcus*, and possibly from other microorganisms through *Enterococcus*, has played an important role in the evolution of *L. casei*.

Lactobacillus casei Intraspecies Diversity

Genomic Islands

One of the major driving forces for bacterial evolution is HGT, whereby bacteriophages, transposons, and other mobile elements are acquired by the host bacterial genome, forming genomic islands (GIs) (Dobrindt et al. 2004). The name GI is derived from the term pathogenicity island, originally coined to describe a cluster of virulence genes identified in *E. coli* (Hacker et al. 1990). Since then, GIs have been noted to contribute to fitness, adaptability, and metabolic versatility of different microorganisms. For example, nitrogen fixation in *Rhizobiaceae* is encoded by "symbiosis islands" (Kaneko et al. 2000, 2002), and "lifestyle adaptation islands" that contain a relatively high number of genes related to carbohydrate utilization are found in *L. plantarum* (Molenaar et al. 2005).

To evaluate the presence of GI in *L. casei*, the genomes of ATCC 334 and BL23 were compared via whole-genome alignment (fig. 3). A total of 12 and 19 GIs (>5 kb), ranging in size from 5.2 kb to 58.2 kb, are identified in ATCC 334 and BL23, respectively (fig. 3 and table 1). Several of the BL23 GIs have been characterized previously, including a gene cluster involved in the catabolism of the cyclic polyol myo-inositol (Yebra et al. 2007). The majority of ATCC 334 GIs encode hypothetical proteins and several encode transcriptional regulators (e.g., TetR and BglG family transcriptional regulators), sugar transporters (e.g., mannose and galactitol PTS systems), and metabolic enzymes (e.g., β-galactosidase and 6-P-α glucosidase). Hypothetical

Table 1
Compositional Features of GIs in *L. casei* ATCC 334

GI	Size (kb)	CDS Coordinates	Mob. CDS[a]	GC %[b]	Function
1	7.1	LSEI_0036–LSEI_0044	2 IS	42.0	Hypothetical protein
2	7.2	LSEI_0084–LSEI_0093	1 IS	39.4	Hypothetical protein
3	9.1	LSEI_0106–LSEI_0114	NF	40.3	Hypothetical protein, transporter, transcriptional regulator
4	58.1	LSEI_0333–LSEI_0388	1 R and 11 IS	45.2	CRISPR, hypothetical protein, ion and amino acid transporters, transcriptional regulator, PTS, acetyltransferase, beta-galactosidase, sugar phosphatase, etc.
5	24.6	LSEI_0564–LSEI_0588	1 R and 4 IS	42.5	Hypothetical protein, type III restriction modification system, levanase, nitroreductase
6	5.3	LSEI_1098–LSEI_1103	3 IS and tRNA	46.5	2-C-methyl-D-erythritol 4-P cytidylyltransferase, glycosyltransferase, hypothetical protein
7	15.9	LSEI_1230–LSEI_1243	2 IS	45.8	Hypothetical protein, transcriptional regulator
8	7.0	LSEI_1506–LSEI_1511	1 IS and tRNA	42.7	Hypothetical protein, transcriptional regulator
9	20.0	LSEI_1938–LSEI_1978	1 I	42.5	Hypothetical protein, prophage, transcriptional regulator
10	14.5	LSEI_2001–LSEI_2011	4 IS	44.1	Hypothetical protein
11	6.7	LSEI_2702–LSEI_2709	NF	44.3	Galactitiol-PTS system, hypothetical protein, transcriptional regulator, triosephosphate isomerase, tagatose-bisphosphate aldolase
12	6.1	LSEI_2767–LSEI_2770	NF	50.0	Hypothetical protein, amidase, phosphohydrolase, alpha-glucosidase

[a] Mob. CDS: Mobility CDS; IS: IS element; R: recombinase; I: integrase; NF: not found.
[b] The GC content of the *L. casei* ATCC 334 chromosome is 46.6%.

proteins account for 55% of the GI-associated genes, approximately twice the percentage identified in the whole genome (27%). Overrepresentation of hypothetical proteins has been observed in nearly all microbial GIs characterized (Hsiao et al. 2005). Nine of the 12 GIs identified in ATCC 334 contain sequences associated with recombination (table 1) (i.e., ISs, recombinases, or integrases). Two of the 12 GIs (GI 6 and GI 8) are located near tRNA, which are common integration sites for mobile elements (Reiter et al. 1989). These results suggest that these regions were recently integrated into the ATCC 334 genome. The majority of the GIs identified differ significantly in GC content from the 46.6% average ATCC 334 chromosomal GC content (table 1), indicating that these sequences are of heterologous origin. The presence of 12 GIs in ATCC 334 suggests that HGT has played a significant role in the evolution, lifestyle adaptation, and metabolic diversity within *L. casei*.

Strain Relationships Determined by MLST

Previously, we developed an MLST scheme and determined the phylogenetic relationship between 40 *L. casei* strains (Cai et al. 2007). The results of that study revealed clusters of strains specific to cheese and silage. Since then, the culture collection has been expanded and now contains 52 *L. casei* strains isolated from different ecological and geographical origins (supplementary table S4, Supplementary Material online). This includes cheeses from three different continents (Australia, Europe, and North America), plant materials (silage, wine, and pickle), and human sources (feces and blood). The single-nucleotide polymorphisms present in five housekeeping gene loci (*ftsZ*, *metRS*, *mutL*, *pgm*, and *polA*) for the 12 new strains were determined and their evolutionary relationships between the 52 strains were evaluated. Overall, MLST separated these strains into 44 *L. casei* sequence types, and the number of alleles ranges from 13 (*metRS* and *polA*) to 21 (*mutL*). A consensus phylogeny using the minimum evolution algorithm resolves two major clusters of strains (supplementary fig. S4, Supplementary Material online). In general, strains of the same origin cluster together, and the significant cluster of cheese isolates observed previously (Cai et al. 2007), designated cluster III in that publication, was confirmed. An analysis of divergence time of different clusters indicates that the divergence of the two major clusters occurred approximately 1.5 million years ago, whereas most cheese isolates diversified approximately 10,000 years ago (supplementary fig. S4, Supplementary Material online). This is consistent with the fact that cheese is a relatively new ecological niche, as cheese manufacture is believed to have begun approximately 8,000 years ago (Fox and McSweeney 2004). To explore the genome diversity and identify the variable genomic regions, 21 strains of *L. casei*, at least one strain from each cluster that diverged 50,000 years ago, was selected for examination by CGH (supplementary fig. S4, Supplementary Material online).

Genome Diversity Revealed by CGH

Analysis by CGH of 21 *L. casei* strains collected from dairy, plant, or human niches revealed that, of 2,661 (97%) chromosomal and 17 (85%) plasmid *L. casei* ATCC 334 CDS surveyed, 1,941 chromosomal CDS (73%) are common to all the strains, representing the common backbone of *L. casei*. This subset includes genes related to central metabolism, replication, transcription, translation, nucleotide metabolism, fatty acid, and phospholipid metabolism. Additionally, all the strains examined contain two (LSEI_0468 and 2270) of the three lactocepin genes, the majority of

Table 2
***Lactobacillus casei* ATCC 334-Specific CDS Revealed by Comparative Genome Hybridization**

GI[c]	Gene ID	Function	GC %[a]
1	LSEI_0036	Hypothetical protein	39
1	LSEI_0038	Hypothetical protein	39
1	LSEI_0039	Hypothetical protein	44
4	LSEI_0344	DNA integration/ recombination/inversion protein	38
4	LSEI_0345	Transposase	44
4	LSEI_0346	Multidrug resistance protein B	41
4	LSEI_0357	Hypothetical protein	53
–	LSEI_0554	Hypothetical protein	41
5	LSEI_0568	Hypothetical protein	44
5	LSEI_0580	Transposase	47
5	LSEI_0586	Type III restriction-modification system methylation subunit (EC 2.1.1.72)	35
8	LSEI_1508	Hypothetical protein	38
8	LSEI_1510	Hypothetical protein	40
8	LSEI_1511	Hypothetical protein	46
–	LSEI_1891	Transposase	45
–	LSEI_1912	Hypothetical protein	47
–	LSEI_1913	Phage protein	48
9	LSEI_1938	Hypothetical protein	35
9	LSEI_1944	Phage protein	44
9	LSEI_1947	Hypothetical protein	46
9	LSEI_1949	Phage protein	47
9	LSEI_1956	Phage protein	43
9	LSEI_1957	Phage-related protein	42
9	LSEI_1965	Hypothetical protein	43
9	LSEI_1966	Hypothetical protein	38
9	LSEI_1967	Hypothetical protein	42
9	LSEI_1968	LexA repressor (EC 3.4.21.88)	43
9	LSEI_1971	Phage protein	39
9	LSEI_1972	Hypothetical protein	35
9	LSEI_1975	Hypothetical protein	35
9	LSEI_1976	Hypothetical protein	35
9	LSEI_1977	Hypothetical protein	33
–	LSEI_A10[b]	Glutamine transport system permease protein GlnP	45
–	LSEI_A11[b]	Glutamine-binding protein	44

[a] GC % of *L. casei* ATCC 334 CDS. The average GC % for ATCC 334 genome is 47%.
[b] Plasmid encoded.
[c] Not found.

Table 3
Number of CDS Missing in Different *L. casei* Strains Examined by Comparative Genome Hybridization

Origin		Strain	Number of CDS[a]	CDS %
Plant	Silage	12A	318	12
	Silage	21/1	269	10
	Silage	32G	282	11
	Wine	A2-309	264	10
	Wine	A2-362	200	7
	Pickle	USDA-P	241	9
	Wine	UCD171	333	12
	Wine	UCD174	302	11
Human	Fecal	DN	240	9
	Fecal	L6	236	9
	Fecal	L9	235	9
	Blood	T7136	235	9
	Blood	T71499	277	10
	Blood	CRF28	302	11
Cheese	Denmark	7A1	406	15
	Denmark	7R1	402	15
	Denmark	83M4	515	19
	Australia	ASCC 1087	507	19
	USA	M36	247	9
	USA	UW1	191	7
	USA	UW4	492	18

[a] Includes both chromosomal and plasmid CDSs.

ATCC 334 CDS predicted to encode peptidases, ~41% ATCC 334 PTS components, and ~64% ATCC 334 transcriptional regulators. Thirty-two chromosomal and two plasmid-coded genes (table 2) are absent in all the test strains, constituting the ATCC 334-specific genes. This group included chromosomal CDS coding for hypothetical proteins, bacteriophages, and transposases as well as plasmid CDS predicted to be involved in glutamine transport. The GC % for the vast majority of these genes deviated widely from the genome average (table 2), suggesting that ATCC 334 recently acquired these CDS. The remaining 701 chromosomal (26%) and 15 plasmid (88%) CDS were divergent in at least one *L. casei* strain tested, which indicated that these CDS constitute the *L. casei* flexible gene pool. Different *L. casei* strains demonstrated variable degrees of absence, ranging from 171 (6%) in strain UW1 to 515 (19%) in 83M4 (table 3).

A hierarchical tree (fig. 4*a*) was constructed based on the overall variability of the CGH data. The most evident cluster in the tree is a group of cheese isolates. Designated as group A cheese isolates (previously referred as cluster III in Cai et al. 2007), this cluster is comprised of *L. casei* UW4, ASCC1087, 7A1, 7R1, and 83M4 and contains the highest number of absent genes. Another group of the cheese isolates, designated group B and comprised of strains UW1 and M36, has among the least number of absent genes. These results indicate that these two groups of cheese isolates have significantly different gene inventories. Five plant isolates (32G, 12A, A2-362, UCD171, and UCD174) demonstrate a stepwise evolutionary pattern and are less closely associated than strains from the group A cheese isolates. The rest of the strains display variable levels of genetic distances relative to other strains. An MLST-based evolutionary tree was also generated for the *L. casei* strains examined by CGH (fig. 4*b*). Comparison of the CGH- and MLST-based phylogeny revealed similar clustering of the group A cheese isolates, whereas a high degree of variation was observed for the rest of the strains. MLST alleles reflect slow evolution of the genome caused by point mutations and selective pressure. In contrast, CGH recognizes large-scale mutation events like insertions and deletions. CGH is thus an indicator of more recent evolutionary changes. Given the overall recombinatorial population structure of *L. casei* (Cai et al. 2007), in which homologous recombination makes the predominant contribution to gene complement differences, variations between CGH- and MLST-based dendrograms are expected.

A total of 25 hypervariable regions in ATCC 334 with respect to other *L. casei* are identified and summarized in figure 5 and table 4. In many cases, the variable regions colocalize with regions of unusual base composition, suggesting recent acquisition of these regions via HGT. Region Y from figure 5 contains CDS from pLSEI1. The majority of pLSEI1 CDS are absent in all *L. casei* strains except for 7A1, suggesting that *L. casei* 7A1 contains a plasmid

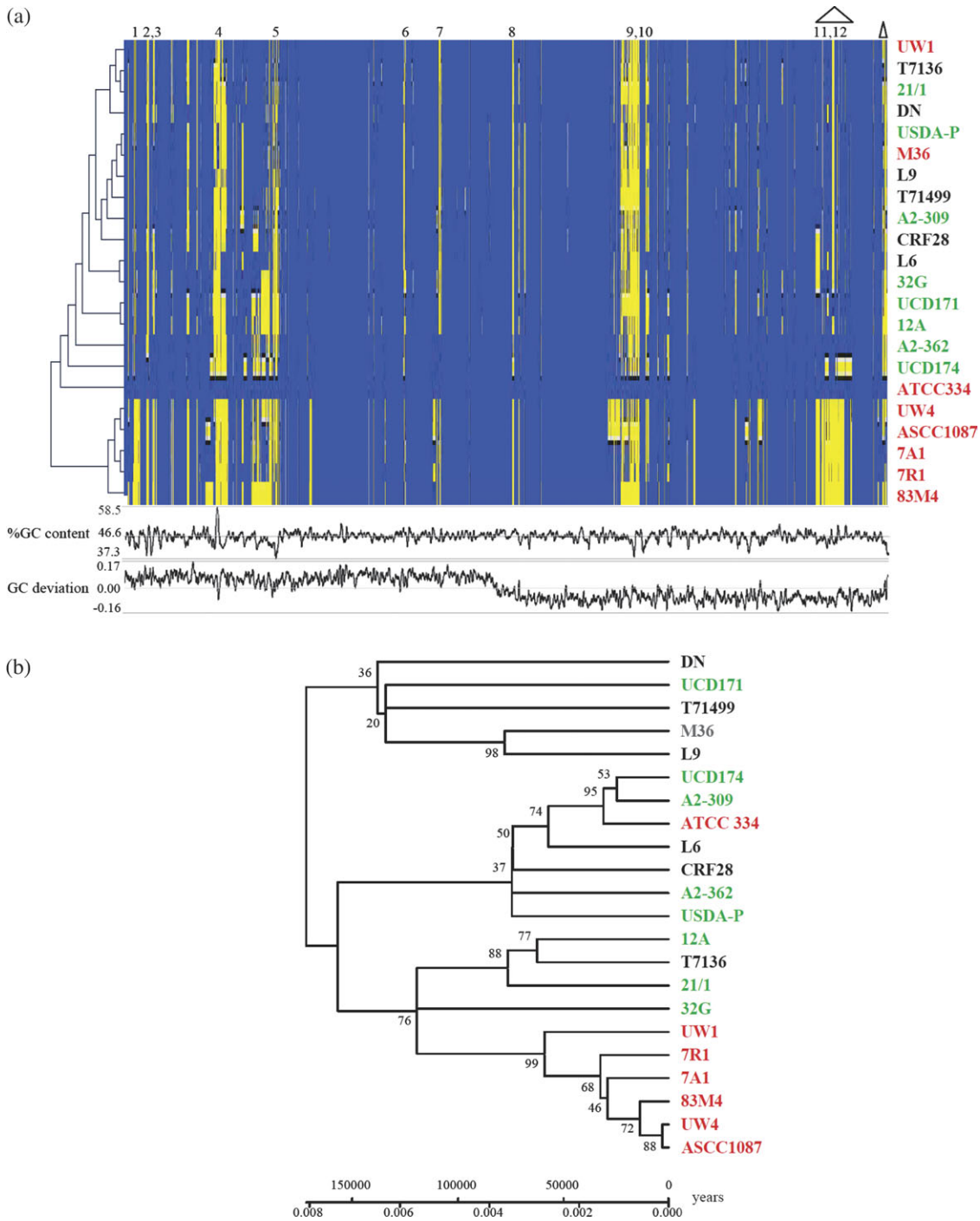


FIG. 4.—Analysis of genome diversity in *L. casei* by CGH and MLST. Panel (*a*) shows CGH composite view of genome diversity among 22 *L. casei* strains isolated from cheeses (red), human sources (black), and plant materials (green). Each row shows the results for one strain, and each column represents the CDS along the ATCC 334 genome, starting at the origin of replication and proceeding clockwise. Blue and yellow areas denote the presence and absence of coding sequences, respectively. Location of the 12 GIs of ATCC 334 (top), the putative lifestyle adaptation island (top, large triangle square), the pLSEI1 (top, small triangle square), the GC %, and GC deviation (bottom) are labeled. Panel (*b*) gives the MLST dendrogram for genetic relatedness among the same *L. casei* strains. The bottom scale shows the divergence time frame and the number of synonymous substitutions per nucleotide site. Bootstrap values on bifurcating branches are based on 1,000 random bootstrap replicates for the consensus tree.

similar to pLSEI1 and that these plasmids are relatively rare within *L. casei*. One main source of these hypervariable regions is GIs. The diversity and distribution of GIs among different *L. casei* strains suggest a large pool of such islands throughout the population. The majorities of these GI-associated hypervariable genes code for hypothetical proteins (57%), phage proteins (9%), and IS elements (8%). Additionally, genes involved in carbohydrate metabolism, amino acid transport, transcriptional regulation, and a type III restriction/modification system are present within GIs.

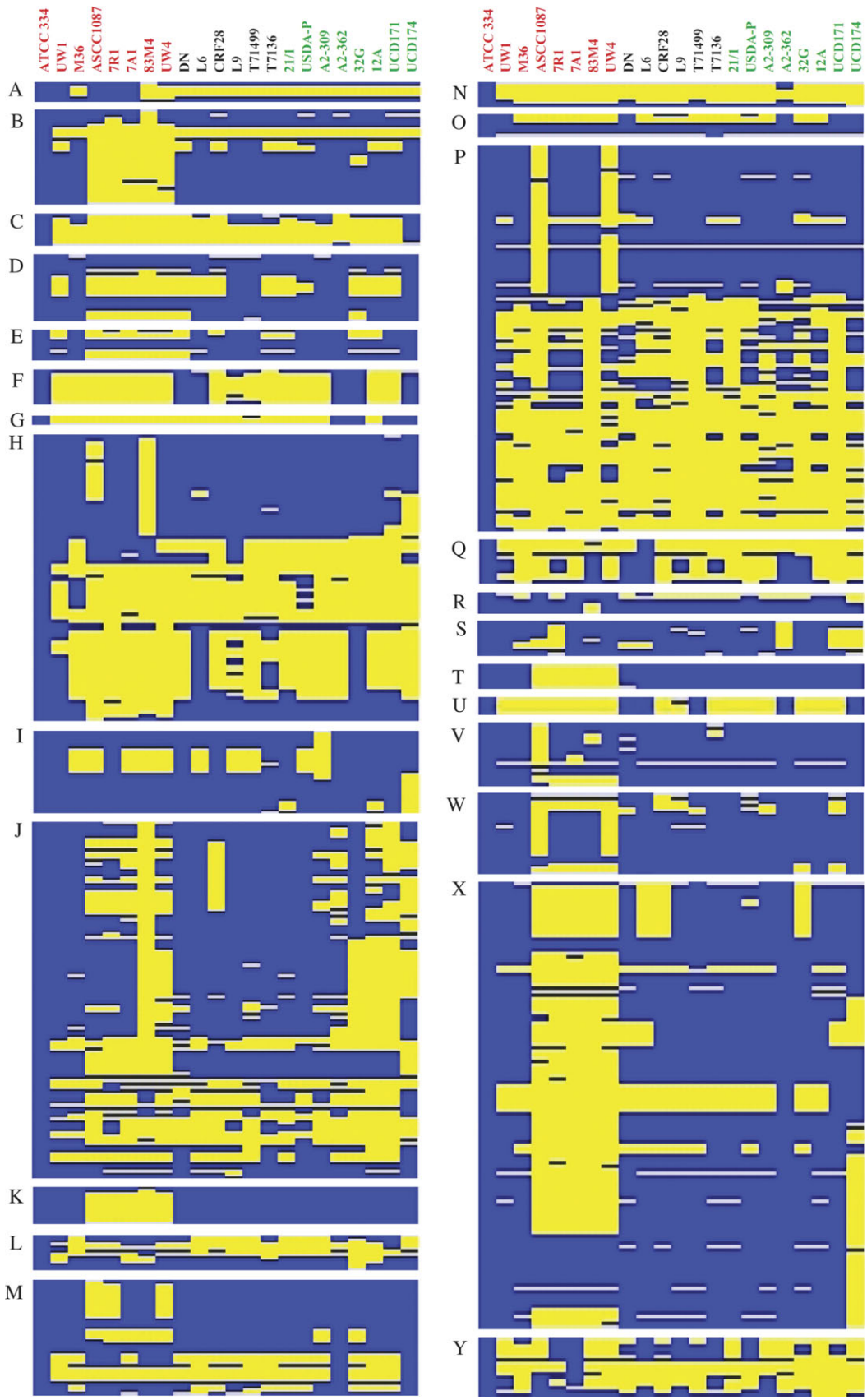


FIG. 5.—Patterns of presence (blue) or absence (yellow) in 25 hypervariable regions of 22 *L. casei* strains isolated from cheeses (red), human sources (black), and plant materials (green). Each row represents a gene, each panel represents a hypervariable region, and each column corresponds to a *L. casei* strain designated vertically across the bottom.

Table 4
Intraspecific Hypervariable Regions in *L. casei*

Region	Gene	GI[b]	Proposed function (other than those encoded by GI)[a]
A	LSEI_0013–LSEI_0017	–	Hypothetical protein
B	LSEI_0031–LSEI_0062	1	Hypothetical protein; dipeptidase; transporter; transcriptional regulator
C	LSEI_0083–LSEI_0093	2	Hypothetical protein
D	LSEI_0102–LSEI_0121	3	Hypothetical protein; transcriptional regulator; transporter; sucrose PTS
E	LSEI_0176–LSEI_0184	–	Hypothetical protein; transcriptional regulator; transporter
F	LSEI_0230–LSEI_0240	–	Hypothetical protein; rhamnosyltransferase; glycosyltransferase; lysozyme
G	LSEI_0270–LSEI_0274	–	Hypothetical protein; amidase; phosphohydrolase
H	LSEI_0302–LSEI_0395	4	Hypothetical protein; regulation, transport, and metabolism of ribose, cellulose PTS, transporter, transcriptional regulator
I	LSEI_0439–LSEI_0465	–	Quinone oxidoreductase; transcriptional regulator; membrane protein; cellobiose PTS; transporter; lactocepin (LSEI_0465)
J	LSEI_0486–LSEI_0600	5	Phage; fructose PTS; thioredoxin; permease; oxidoreductase; lysozyme
K	LSEI_0708–LSEI_0717	–	Glycosyltransferase; transcriptional regulator; oxidoreductase; quinine reductase
L	LSEI_1094–LSEI_1107	6	Endopeptidase; metabolism and transport of beta-glucoside; transcriptional regulator
M	LSEI_1208–LSEI_1243	7	Transcriptional regulator; transporter; hypothetical protein; acetyltransferase
N	LSEI_1506–LSEI_1513	8	Hypothetical protein
O	LSEI_1529–LSEI_1535	–	Hypothetical protein; quinone oxidoreductase; transcriptional regulator
P	LSEI_1862–LSEI_1978	9	Phage; transcriptional regulator; hypothetical protein; oligopeptide transporter
Q	LSEI_2001–LSEI_2018	10	Cell wall/membrane biosynthesis
R	LSEI_2050–LSEI_2058	–	Hypothetical protein; transcriptional regulator; acyltransferase; short-chain dehydrogenase
S	LSEI_2087–LSEI_2100	–	Hypothetical protein; type I R/M system
T	LSEI_2191–LSEI_2197	–	Hypothetical protein; ABC transporter; 6-P-beta-glucosidase; beta-glucoside PTS; transcriptional regulator
U	LSEI_2341–LSEI_2351	–	Phosphopentomutase; transporter; uridine phosphorylase; carbohydrate diacid regulator
V	LSEI_2386–LSEI_2406	–	Hypothetical protein; protease; transcriptional regulator
W	LSEI_2435–LSEI_2457	–	Hypothetical protein; phosphonates transport; ABC transporter; acyl-CoA synthetase; acyltransferase
X	LSEI_2658–LSEI_2792	11, 12	Hypothetical protein; PTS systems for mannose, *N*-acetylgalactosamine, glucose, 3-keto-L-gulonate, and fructose; transcriptional regulator; sugar transport and metabolism; acetyltransferase
Y	LSEI_A01–LSEI_A20	–	pLSEI1

[a] Transposases are not listed.
[b] Not found.

Just as pathogenicity islands alter the host specificity and virulence of pathogenic bacteria, acquisition of GIs may draw novel genes required to adapt to new ecological niches from the environment. Characterization of these newly acquired genes may allow us to gain new insights into the physiological and molecular toolbox necessary for the adaptation and evolution of *L. casei* to different ecological niches.

Among the 25 hypervariable regions identified, region X is proposed as a *L. casei* putative lifestyle adaptation island. Close to the origin of replication, region X runs from 2,643,036 bp to 2,771,902 bp (LSEI_2658–2792), amounting to 128.9 kb and 4.5% of ATCC 334 chromosome. It shows strong overrepresentation of genes involved in carbohydrate metabolism (fig. 1), representing ~16% of all carbohydrate-related genes in the genome. Predicted gene functions in this region include: complete or incomplete PTS systems for mannose, *N*-acetylgalactosamine, glucose, galactitol, sorbitol, 3-keto-L-gulonate, and fructose; other carbohydrate transporters with unknown substrates; a two-component transcriptional regulator (LSEI_2680 and 2681) and 19 other transcriptional regulators; and a wide variety of metabolic enzymes including alpha-glucosidase, alpha-L-fucosidase, amidohydrolase, beta-lactamase, hyaluronate lyase, and unsaturated glucuronyl hydrolase. It also contains genes predicted to be involved in amino acid metabolism and pentose and glucuronate interconversion pathways as well as seven acetyltransferase-encoding genes, genes coding for oxidoreductases, and a high number of genes with unknown function. Unlike the lifestyle adaptation regions in *L. plantarum* (Molenaar et al. 2005), region X does not display unusual base composition. This suggests that region X was either acquired from an organism with similar base composition or has undergone reductive evolution where gene loss is the driving force for the plasticity of the region.

The tight clustering of the group A cheese isolates in both the MLST and CGH analyses as well as the relatively high number (>120) of genes absent in this group compared with the other *L. casei* strains tested (figs. 4*a* and 5 regions B, K, M, T, and X; table 3; and supplementary table S5, Supplementary Material online) indicate that this group is genetically distinct from other *L. casei* strains. Genes absent from this group include CDS predicted to be involved in carbohydrate utilization and transcriptional regulation. Of particular interest is the finding that ~76% of the putative lifestyle adaptation island genes, including all carbohydrate-related genes except for the mannose PTS system, are deleted in this group. Concurrence of such large-scale deletions is remarkable, given the different geographical locations from which these *L. casei* strains were isolated. Such

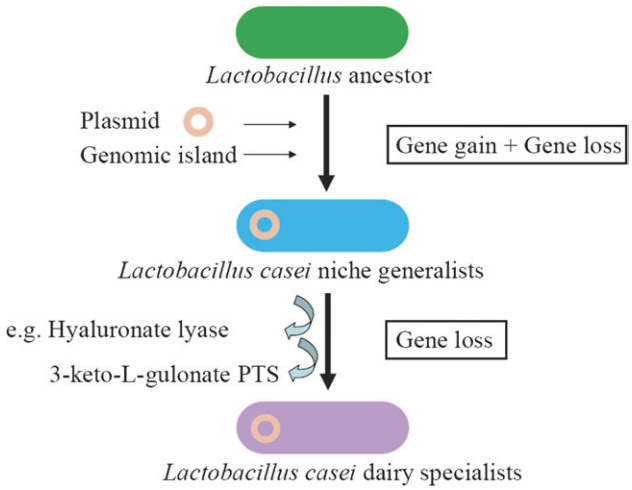


FIG. 6.—Model of evolution of *L. casei* from an ancestral *Lactobacillus*. Evolution is achieved by both gene gain and gene loss.

deletions suggest that the group A cheese isolates have evolved toward becoming dairy specialists by gene decay, during which metabolic pathways not required by the dairy niche are passively lost. High numbers of IS elements present in *L. casei* may have facilitated this process.

In contrast, group B cheese isolates, and possibly other *L. casei* strains examined, seem to be niche generalists. Unlike niche specialists, niche generalists exploit a variety of habitats and experience a wide range of environmental conditions. Consequently, they require genes to detect, assess, and handle the range of physicochemical conditions and environmental stresses present in the variety of environments they may encounter. This explains why group B cheese isolates contain less variable CDS compared with the group A cheese isolates (table 3). The versatile adaptability of niche generalists also makes it difficult to match their clustering patterns with their origin of isolation.

Conclusion

Lactobacillus casei is considered a versatile species that has evolved to occupy diverse habitats for hundreds of millions of years. These habitats include raw and fermented dairy products, raw and fermented plant materials, and reproductive and gastrointestinal tracts of humans and animals. The plant- and vertebrate-associated niches of *L. casei* are dynamic, nutritionally variable, and ancient; whereas the dairy niche is constant, nutrient rich, and relatively recent, with cheese manufacturing believed to have begun approximately 8,000 years ago (Fox and McSweeney 2004). Availability of genome sequence of *L. casei* ATCC 334, a cheese isolate, allows for an analysis of its metabolic diversity and genetic "fitness" to inhabit various ecological niches. When compared with other sequenced lactobacilli, *L. casei* ATCC 334 contains a relatively high number of genes involved in carbohydrate metabolism as well as transcriptional regulation and signal transduction. These results are consistent with ATCC 334 being capable of inhabiting a wide variety of ecological niches. Comparison of the ATCC 334 genome with the *L. casei* BL23 genome and CGH results allowed for the identification of GIs and hypervariable regions, including a putative lifestyle adaptation island. The genes present in the GIs and hypervariable regions include a high number of genes involved in carbohydrate metabolism and regulation, suggesting that HGT played a significant role in the adaptation of *L. casei* to novel niches by acquisition of foreign genes. These genes were acquired from a diverse group of non *Lactobacillus*/*Pediococcus* Gram-positive microorganisms; however, the most common source was *Entercococcus*, suggesting that this genus, likely due to its promiscuous nature, has played a significant role in the exchange of genetic material within LAB. In addition to HGT, gene duplication followed by mutation and vertical inheritance has played a significant role in the evolution of *L. casei*. This is particularly evident in the proteolytic enzyme (PepC/E, PepO, and lactocepins) and lactate dehydrogenase systems. Finally, the CGH results clearly demonstrate that the *L. casei* species contains a distinct subpopulation, designated the group A cheese isolates, that have undergone significant gene decay (fig. 6). This subpopulation is missing more than 120 CDS (supplementary table S5, Supplementary Material online), the majority of which are associated with transcriptional regulation and carbohydrate utilization (e.g., hyaluronate lyase and 3-keto-L-gulonate utilization). Loss of these genes likely has led to organisms with increased fitness in the dairy niche but decreased capacity to inhabit other ecological niches; hence, they are thought of as dairy specialists. At some point in the future, these dairy specialists may become a separate species. In summary, all three processes known to be involved in the evolution of microorganism, modification of existing genes via mutation followed by vertical inheritance, gain of genes via HGT, and loss of genes no longer necessary have contributed to the evolution of *L. casei*.

Supplementary Material

Supplementary figures S1–S4 and tables S1–S5 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Funding

This work was supported by Danisco Inc., Dairy Management, Inc. through the Center for Dairy Research, the College of Agricultural and Life Sciences at the University of Wisconsin, and the United States Department of Agriculture.

Acknowledgments

We thank David Mills (Department of Viticulture and Enology, University of California at Davis, CA) and Minna Salminen (Department of Medicine, Helsinki University Central Hospital, Helsinki, Finland) for providing strains of *L. casei*. We thank Theresa Walunas and Integrated Genomics (Chicago, IL) for assistance with the microarray design. We thank Philippe Horvath (Danisco France SAS) and Rodolphe Barrangou (Danisco USA Inc.) for critical

reading of the manuscript. We thank Kanokwan Tandee (Department of Food Science, University of Wisconsin, Madison, WI) for discussion on carbohydrate utilization of *L. casei*. We thank Anna Cisler (Department of Genetics, University of Wisconsin, Madison, WI) for help with MLST sequencing. Peggy Steele, a member of Dr Steele's family, is employed by Danisco Inc., a supplier of bacterial cultures to the food industry.

Literature Cited

Abbott JC, Aanensen DM, Rutherford K, Butcher S, Spratt BG. 2005. WebACT—an online companion for the Artemis Comparison Tool. Bioinformatics. 21:3665–3666.

Acedo-Felix E, Perez-Martinez G. 2003. Significant differences between *Lactobacillus casei* subsp. *casei* ATCC 393T and a commonly used plasmid-cured derivative revealed by a polyphasic study. Int J Syst Evol Microbiol. 53:67–75.

Altermann E, et al. 2005. Complete genome sequence of the probiotic lactic acid bacterium *Lactobacillus acidophilus* NCFM. Proc Natl Acad Sci USA. 102:3906–3912.

Axelsson L. 2004. Lactic acid bacteria: classification and physiology. In: Salminen S, von Wright A, Ouwehand A, editors. Lactic acid bacteria microbiological and functional aspects, 3rd edition. New York: Marcel Dekker Inc. p. 1–66.

Azcarate-Peril MA, et al. 2008. Analysis of the genome sequence of *Lactobacillus gasseri* ATCC 33323 reveals the molecular basis of an autochthonous intestinal organism. Appl Environ Microbiol. 74:4610–4625.

Barrangou R, et al. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 315:1709–1712.

Barre O, Mourlane F, Solioz M. 2007. Copper induction of lactate oxidase of *Lactococcus lactis*: a novel metal stress response. J Bacteriol. 189:5947–5954.

Bolotin A, et al. 2004. Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus*. Nat Biotechnol. 22:1554–1558.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 19:185–193.

Cai H, Rodríguez BT, Zhang W, Broadbent JR, Steele JL. 2007. Genotypic and phenotypic characterization of *Lactobacillus casei* strains isolated from different ecological niches suggests frequent recombination and niche specificity. Microbiology. 153:2655–2665.

Callanan M, et al. 2008. Genome sequence of *Lactobacillus helveticus*, an organism distinguished by selective gene loss and insertion sequence element expansion. J Bacteriol. 190:727–735.

Chen YS, Christensen JE, Broadbent JR, Steele JL. 2003. Identification and characterization of *Lactobacillus helveticus* PepO2, an endopeptidase with post-proline specificity. Appl Environ Microbiol. 69:1276–1282.

Christensen JE, Dudley EG, Pederson JA, Steele JL. 1999. Peptidases and amino acid catabolism in lactic acid bacteria. Antonie Van Leeuwenhoek. 76:217–246.

Christiansen JK, et al. 2008. Phenotypic and genotypic analysis of amino acid auxotrophy in *Lactobacillus helveticus* CNRZ 32. Appl Environ Microbiol. 74:416–423.

Cole ST, et al. 2001. Massive gene decay in the leprosy bacillus. Nature. 409:1007–1011.

De Koning AP, Brinkman FS, Jones SJ, Keeling PJ. 2000. Lateral gene transfer and metabolic adaptation in the human parasite *Trichomonas vaginalis*. Mol Biol Evol. 17:1769–1773.

Devriese LA, Collins MD, Wirth R. 1992. The genus *Enterococcus*. In: Balows A, Trüper HG, Dworkin M, Harder W, Schleifer KH, editors. The Prokaryotes: A handbook on the biology of bacteria: ecophysiology, isolation, identification, applications, 2nd edition. New York: Springer. p. 1465–1481.

De Vuyst L, Degeest B. 1999. Heteropolysaccharides from lactic acid bacteria. FEMS Microbiol Rev. 23:153–177.

De Vuyst L, Leroy F. 2007. Bacteriocins from lactic acid bacteria: production, purification, and food applications. J Mol Microbiol Biotechnol. 13:194–199.

Dobrindt U, Hochhut B, Hentschel U, Hacker J. 2004. Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol. 2:414–424.

Doolittle RF, Feng DF, Tsang S, Cho G, Little E. 1996. Determining divergence times of the major kingdoms of living organisms with a protein clock. Science. 271:470–477.

Elena SF, Lenski RE. 2003. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat Rev Genet. 4:457–469.

FAO/WHO. 2002. Food and agriculture organization of united Nation and world health organization working group report on drafting guidelines for the evaluation of probiotics in food. London: FAO.

Feldgarden M, Byrd N, Cohan FM. 2003. Gradual evolution in bacteria: evidence from *Bacillus* systematics. Microbiology. 149:3565–3573.

Fox PF, McSweeney PLH. 2004. Cheese: an overview. In: Fox PF, McSweeney PLH, Cogan TM, Guinee TP, editors. Cheese Chemistry, Physics and Microbiology, 3rd edition. San Diego (CA): Elsevier. p. 1–37.

Garvie EI. 1980. Bacterial lactate dehydrogenases. Microbiol Rev. 44:106–139.

Giraud A, et al. 2001. Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. Science. 291:2606–2608.

Haandrikman AJ, Kok J, Venema G. 1991. Lactococcal proteinase maturation protein PrtM is a lipoprotein. J Bacteriol. 173:4517–4525.

Hacker J, et al. 1990. Deletions of chromosomal regions coding for fimbriae and hemolysins occur *in vitro* and *in vivo* in various extraintestinal *Escherichia coli* isolates. Microb Pathog. 8:213–225.

Haft DH, Selengut J, Mongodin EF, Nelson KE. 2005. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput Biol. 1:e60.

Hoefnagel MH, et al. 2002. Metabolic engineering of lactic acid bacteria, the combined approach: kinetic modelling, metabolic control and experimental analysis. Microbiology. 148: 1003–1013.

Hols P, et al. 2005. New insights in the molecular biology and physiology of *Streptococcus thermophilus* revealed by comparative genomics. FEMS Microbiol Rev. 29:435–463.

Horvath P, et al. 2008. Comparative analysis of CRISPR loci in lactic acid bacteria genomes. Int J Food Microbiol. 131:62–70.

Hsiao WW, Ung K, Aeschliman D, Bryan J, Finlay BB, Brinkman FS. 2005. Evidence of a large novel gene pool associated with prokaryotic genomic islands. PLoS Genet. 1:e62.

Huson DH. 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics. 14:68–73.

Izutsu K, et al. 2008. Comparative genomic analysis using microarray demonstrates a strong correlation between the presence of the 80-kilobase pathogenicity island and pathogenicity in Kanagawa phenomenon-positive *Vibrio parahaemolyticus* strains. Infect Immun. 76:1016–1023.

Kandler O. 1983. Carbohydrate metabolism in lactic acid bacteria. Antonie Van Leeuwenhoek. 49:209–224.

Kandler O, Weiss N. 1986. Genus *Lactobacillus*. In: Sneath PHA, Mair NS, Sharpe ME, Holt JG, editors. Bergey's manual of systematic bacteriology, 9th edition. Baltimore (MD): Williams & Wilkins. p. 1063–1065.

Kaneko T, et al. 2000. Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. DNA Res. 7:331–338.

Kaneko T, et al. 2002. Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. DNA Res. 9:189–197.

Kleerebezem M, et al. 2003. Complete genome sequence of *Lactobacillus plantarum* WCFS1. Proc Natl Acad Sci USA. 100:1990–1995.

Kumar S, Tamura K, Nei M. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform. 5:150–163.

Lawrence JG. 1999. Gene transfer, speciation, and the evolution of bacterial genomes. Curr Opin Microbiol. 2:519–523.

Lawrence JG, Ochman H. 1998. Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci USA. 95:9413–9417.

Mack DR, Ahrne S, Hyde L, Wei S, Hollingsworth MA. 2003. Extracellular MUC3 mucin secretion follows adherence of *Lactobacillus* strains to intestinal epithelial cells *in vitro*. Gut. 52:827–833.

Makarova K, et al. 2006. Comparative genomics of the lactic acid bacteria. Proc Natl Acad Sci USA. 103:15611–15616.

Marco ML, Pavan S, Kleerebezem M. 2006. Towards understanding molecular modes of probiotic action. Curr Opin Biotechnol. 17:204–210.

Mata L, Erra-Pujada M, Gripon JC, Mistou MY. 1997. Experimental evidence for the essential role of the C-terminal residue in the strict aminopeptidase activity of the thiol aminopeptidase PepC, a bacterial bleomycin hydrolase. Biochem J. 328(Pt 2):343–347.

Mata L, Gripon JC, Mistou MY. 1999. Deletion of the four C-terminal residues of PepC converts an aminopeptidase into an oligopeptidase. Protein Eng. 12:681–686.

Maurelli AT, Fernandez RE, Bloch CA, Rode CK, Fasano A. 1998. "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. Proc Natl Acad Sci USA. 95:3943–3948.

Mayra-Makinen A, Bigret M. 1998. Industrial use and production of lactic acid bacteria. In: Salminen S, Wright AV, editors. Lactic acid bacteria—microbiology and functional aspects, 2nd edition. New York: Marcel Dekker Inc. p. 73–102.

McKay LL. 1983. Functional properties of plasmids in lactic streptococci. Antonie Van Leeuwenhoek. 49:259–274.

McLandsborough LA, Kolaetis KM, Requena T, McKay LL. 1995. Cloning and characterization of the abortive infection genetic determinant *abiD* isolated from pBF61 of *Lactococcus lactis* subsp. *lactis* KR5. Appl Environ Microbiol. 61:2023–2026.

McLysaght A, Baldi PF, Gaut BS. 2003. Extensive gene gain associated with adaptive evolution of poxviruses. Proc Natl Acad Sci USA. 100:15655–15660.

Mirkin BG, Fenner TI, Galperin MY, Koonin EV. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol. 3:e2.

Molenaar D, et al. 2005. Exploring *Lactobacillus plantarum* genome diversity by using microarrays. J Bacteriol. 187:6119–6127.

Monedero V, et al. 2007. The phosphotransferase system of *Lactobacillus casei*: regulation of carbon metabolism and connection to cold shock response. J Mol Microbiol Biotechnol. 12:20–32.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature. 405:299–304.

Ogata H, et al. 2001. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science. 293:2093–2098.

Ohta T. 2003. Evolution by gene duplication revisited: differentiation of regulatory elements versus proteins. Genetica. 118:209–216.

Oppegard C, et al. 2007. The two-peptide class II bacteriocins: structure, production, and mode of action. J Mol Microbiol Biotechnol. 13:210–219.

Orsi RH, Ripoll DR, Yeung M, Nightingale KK, Wiedmann M. 2007. Recombination and positive selection contribute to evolution of *Listeria monocytogenes inlA*. Microbiology. 153:2666–2678.

Paulsen IT, et al. 2003. Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. Science. 299: 2071–2074.

Reiter WD, Palm P, Yeats S. 1989. Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. Nucleic Acids Res. 17:1907–1914.

Rico J, Yebra MJ, Perez-Martinez G, Deutscher J, Monedero V. 2008. Analysis of *ldh* genes in *Lactobacillus casei* BL23: role on lactic acid production. J Ind Microbiol Biotechnol. 35:579–586.

Riley MA, Wertz JE. 2002. Bacteriocins: evolution, ecology, and application. Annu Rev Microbiol. 56:117–137.

Ruas-Madiedo P, Gueimonde M, Margolles A, de los Reyes-Gavilan CG, Salminen S. 2006. Exopolysaccharides produced by probiotic strains modify the adhesion of probiotics and enteropathogens to human intestinal mucus. J Food Prot. 69:2011–2015.

Saeed AI, et al. 2003. TM4: a free, open-source system for microarray data management and analysis. Biotechniques. 34:374–378.

Saito A, Fujii T, Miyashita K. 2003. Distribution and evolution of chitinase genes in *Streptomyces* species: involvement of gene-duplication and domain-deletion. Antonie Van Leeuwenhoek. 84:7–15.

Salminen MK, et al. 2006. *Lactobacillus* bacteremia, species identification, and antimicrobial susceptibility of 85 blood isolates. Clin Infect Dis. 42:e35–e44.

Savijoki K, Ingmer H, Varmanen P. 2006. Proteolytic systems of lactic acid bacteria. Appl Microbiol Biotechnol. 71:394–406.

Schneider D, Lenski RE. 2004. Dynamics of insertion sequence elements during experimental evolution of bacteria. Res Microbiol. 155:319–327.

Smeianov VV, et al. 2007. Comparative high-density microarray analysis of gene expression during growth of *Lactobacillus helveticus* in milk versus rich culture medium. Appl Environ Microbiol. 73:2661–2672.

Sokurenko EV, et al. 1998. Pathogenic adaptation of *Escherichia coli* by natural variation of the FimH adhesin. Proc Natl Acad Sci USA. 95:8922–8926.

Sorek R, Kunin V, Hugenholtz P. 2008. CRISPR–a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat Rev Microbiol. 6:181–186.

Springael D, Top EM. 2004. Horizontal gene transfer and microbial adaptation to xenobiotics: new types of mobile genetic elements and lessons from ecological studies. Trends Microbiol. 12:53–58.

Sridhar VR, Hughes JE, Welker DL, Broadbent JR, Steele JL. 2005. Identification of endopeptidase genes from the genomic

sequence of *Lactobacillus helveticus* CNRZ32 and the role of these genes in hydrolysis of model bitter peptides. Appl Environ Microbiol. 71:3025–3032.

Tanaka K, et al. 2002. Two different pathways for D-xylose metabolism and the effect of xylose concentration on the yield coefficient of L-lactate in mixed-acid fermentation by the lactic acid bacterium *Lactococcus lactis* IO-1. Appl Microbiol Biotechnol. 60:160–167.

Tenaillon O, Taddei F, Radmian M, Matic I. 2001. Second-order selection in bacterial evolution: selection acting on mutation and recombination rates in the course of adaptation. Res Microbiol. 152:11–16.

Top EM, Springael D. 2003. The role of mobile genetic elements in bacterial adaptation to xenobiotic organic compounds. Curr Opin Biotechnol. 14:262–269.

Tuomola EM, Salminen SJ. 1998. Adhesion of some probiotic and dairy *Lactobacillus* strains to Caco-2 cell cultures. Int J Food Microbiol. 41:45–51.

van de Guchte M, et al. 2006. The complete genome sequence of *Lactobacillus bulgaricus* reveals extensive and ongoing reductive evolution. Proc Natl Acad Sci U S A. 103:9274–9279.

Ventura M, et al. 2006. Comparative genomics and transcriptional analysis of prophages identified in the genomes of *Lactobacillus gasseri*, *Lactobacillus salivarius*, and *Lactobacillus casei*. Appl Environ Microbiol. 72:3130–3146.

Viana R, Yebra MJ, Galan JL, Monedero V, Perez-Martinez G. 2005. Pleiotropic effects of lactate dehydrogenase inactivation in *Lactobacillus casei*. Res Microbiol. 156:641–649.

Yebra MJ, et al. 2007. Identification of a gene cluster enabling *Lactobacillus casei* BL23 to utilize myo-inositol. Appl Environ Microbiol. 73:3850–3858.

Yu W, Gillies K, Kondo JK, Broadbent JR, McKay LL. 1996. Loss of plasmid-mediated oligopeptide transport system in lactococci: another reason for slow milk coagulation. Plasmid. 35:145–155.

William Martin, Associate Editor

Accepted July 10, 2009