Genotypic and phenotypic characterization of *Lactobacillus casei* strains isolated from different ecological niches suggests frequent recombination and niche specificity

Hui Cai,¹ Beatriz T. Rodríguez,² Wei Zhang,³ Jeff R. Broadbent² and James L. Steele¹

Correspondence
James L. Steele
jlsteele@wisc.edu

¹Department of Food Science, 1605 Linden Dr., University of Wisconsin, Madison, WI 53706, USA

²Utah Veterinary Diagnostic Laboratory, 950 East 1400 North, Logan, UT 84322, USA

³National Center for Food Safety and Technology, Illinois Institute of Technology, Summit, IL 60501, USA

Lactobacillus casei strains are lactic acid bacteria (LAB) that colonize diverse ecological niches, and have broad commercial applications. To probe their evolution and phylogeny, 40 *L. casei* strains were characterized; the strains included isolates from plant materials ($n=9$), human gastrointestinal tracts ($n=7$), human blood ($n=1$), cheeses from different geographical locations ($n=22$), and one strain of unknown origin. API biochemical testing identified niche-specific carbohydrate fermentation profiles. A multilocus sequence typing (MLST) scheme was developed for *L. casei*. Partial sequencing of six housekeeping genes (*ftsZ*, *metRS*, *mutL*, *nrdD*, *pgm* and *polA*) revealed between 11 (*nrdD*) and 20 (*mutL*) allelic types, as well as 36 sequence types. Phylogenetic analysis of MLST data by Reticulate and split decomposition analysis indicated frequent intra-species recombination. Purifying selection was detected, and is likely to have contributed to the evolution of certain *L. casei* genes. Pulsed-field gel electrophoresis (PFGE) using *SfiI* was able to discriminate all the isolates, even those not differentiated by MLST. Phylogenetic trees reconstructed based on the MLST data using minimum evolution algorithm, and the *SfiI*-PFGE restriction patterns using the unweighted-pair group method with arithmetic mean (UPGMA), revealed consensus clusters of strains specific to cheese and silage. Topological discrepancies between the MLST and PFGE trees were also observed, suggesting that intragenic point mutations have accumulated at a slower rate than indels and genome rearrangements in *L. casei*. The *L. casei* population analysed in this study demonstrated both a high level of phenotypic and genotypic diversity, as well as specificity to different ecological niches.

Received 26 January 2007

Revised 5 April 2007

Accepted 13 April 2007

INTRODUCTION

Lactobacillus casei strains are Gram-positive, facultatively anaerobic, industrially important lactic acid bacteria (LAB) that have been primarily used as probiotics and speciality

Abbreviations: DI, discrimination index; d_N , number of non-synonymous substitutions per non-synonymous site; d_S , number of synonymous substitutions per synonymous site; GI, gastrointestinal; HGT, horizontal gene transfer; LAB, lactic acid bacteria; ME, minimum evolution; MLST, multilocus sequence typing; PFGE, pulsed-field gel electrophoresis; ST, sequence type; SNP, single nucleotide polymorphism; UPGMA, unweighted-pair group method with arithmetic mean.

The GenBank/EMBL/DDBJ accession numbers for the sequences reported in this paper are EF538428–EF538467 (*ftsZ*), EF538468–EF538507 (*metRS*), EF538508–EF538547 (*mutL*), EF538548–EF538587 (*nrdD*), EF538588–EF538627 (*pgm*) and EF538628–EF538667 (*polA*).

cultures for cheese flavour development (Mayra-Makinen & Bigret, 1998). Their broad commercial applications may reflect their remarkable ecological adaptability to diverse habitats. *L. casei* may be isolated from raw and fermented dairy products, intestinal tracts and reproductive systems of humans and animals, as well as fresh and fermented plant products (Kandler & Weiss, 1986). The genetic basis for ecological flexibility in *L. casei* is not fully understood; however, comparative genomic analyses have suggested extensive gene loss and gene acquisitions during evolution of lactobacilli, presumably via bacteriophage- or conjugation-mediated horizontal gene transfers (HGTs), and these may have facilitated their adaptation to diverse ecological niches (Makarova *et al.*, 2006). For example, milk- and vegetable-associated subspecies of *Lactobacillus delbrueckii* have a high level of genetic heterogeneity, and correlations have been shown between specific gene loss/acquisition and the

ability of this species to colonize specific habitats (Germond *et al.*, 2003). Moreover, comparative genomic analysis on 20 *Lactobacillus plantarum* strains of various sources revealed genomic regions with unusual base composition, indicative of evolutionarily recent acquisitions (Molenaar *et al.*, 2005).

Molecular typing of *L. casei* is crucial to understanding the evolutionary adaptation of this species to different ecological niches. Moreover, definitive identification of *L. casei* at the strain level is important for a variety of industrial applications, as it facilitates tracking of specific strains with industrially relevant properties, such as probiotic, sensorial or antimicrobial attributes. To date, several molecular typing approaches, including pulsed-field gel electrophoresis (PFGE; Tynkkynen *et al.*, 1999), randomly amplified polymorphic DNA (Tynkkynen *et al.*, 1999), rRNA restriction fragment length polymorphism (Chen *et al.*, 2000), temporal temperature-gradient gel electrophoresis (Vasquez *et al.*, 2001), and repetitive element PCR (Michael *et al.*, 2006), have been applied to *L. casei*, with PFGE reported to provide the highest discriminatory power among these methods. However, these techniques have less utility in defining underlying phylogenetic relationships, and multilocus sequence typing (MLST) is of value in this regard (Enright & Spratt, 1999). By partially sequencing six or seven housekeeping genes, MLST characterizes the alleles present at several relatively conserved genomic loci and, as a result, differentiates bacterial strains. First introduced in 1998 (Maiden *et al.*, 1998), MLST has been used to characterize many bacterial pathogens (Lacher *et al.*, 2007; Olvera *et al.*, 2006; Nightingale *et al.*, 2005) and several LAB species, such as *Oenococcus oeni* (de las Rivas *et al.*, 2004) and *L. plantarum* (de las Rivas *et al.*, 2006), but it has not yet been applied to *L. casei*. Additionally, bacterial population structures can often be inferred from the MLST data. While the population structures for bacterial pathogens are often found to be clonal (Olvera *et al.*, 2006) or epidemic (Miragaia *et al.*, 2007), recent MLST studies of two LAB species, *O. oeni* and *L. plantarum*, have demonstrated that both species have panmictic non-clonal population structures, suggesting substantial recombination (de las Rivas *et al.*, 2004, 2006).

The goals of this study were to gain comprehensive knowledge of the phenotypic and genotypic characteristics of *L. casei* isolated from different environments [cheeses, fermented plant materials, human gastrointestinal (GI) tracts and human blood] and a better understanding of the evolutionary adaptation of *L. casei* to different ecological niches. To achieve this goal, we assembled a set of 40 *L. casei* isolates from various sources, and used these strains to: (i) develop an MLST scheme for *L. casei*; (ii) apply MLST to assess phylogenetic relationship and evolutionary characteristics of these isolates; (iii) identify niche-specific phenotypic and genotypic traits; and (iv) compare, at a methodological level, the discriminatory powers of MLST and PFGE for *L. casei*.

METHODS

Bacterial strains. A total of 40 *L. casei* strains were selected and characterized in this study (Table 1). These included strains isolated from fermented plant materials ($n=9$), human GI tracts ($n=7$), a human blood sample from an immunocompromised patient ($n=1$), cheeses from different geographical locations ($n=22$), and one strain of unknown origin. Stock cultures were stored at -80°C in 20% (v/v) glycerol. Working cultures were prepared from frozen stock by two transfers in MRS broth (BD Biosciences), without shaking, for 16–18 h at 37°C .

API biochemical testing. API tests were performed as described previously (Broadbent *et al.*, 2003), except that *L. casei* strains were incubated at 37°C . API results of 3, 4 and 5 were interpreted as positive, whereas 0, 1 and 2 were interpreted as negative. When calculating percentage frequencies of strains able to utilize carbohydrates, 1 was given for positive results, and 0 was given for negative results.

PFGE. PFGE gel plugs were prepared utilizing the CHEF Genomic DNA Plug Kits for bacterial DNA (Bio-Rad). Agarose-embedded DNA was digested with 50 U *Sfi*I (Promega) for 16–18 h at 50°C . The restriction fragments were separated by electrophoresis in a 1% PFGE certified agarose (Bio-Rad), using a CHEF DR II apparatus (Bio-Rad) in $0.5\times$ Tris borate EDTA buffer as follows: initial switch time, 1.0 s; final switch time, 20.0 s; start ratio, 1.0; temperature, 14°C ; run time, 22 h; voltage, 200 V. The gels were stained in ethidium bromide solution (10 mg ml^{-1}) for 20 min, followed by three distilled water washes. DNA fingerprint patterns were interpreted by Bionumerics 4.0 software (Applied Maths). A dendrogram representing strain relatedness was determined using the unweighted pair group method using arithmetic means (UPGMA) with Dice coefficients based on the *Sfi*I restriction profiles for PFGE.

MLST loci selection. Intragenic regions of six housekeeping genes were selected for the MLST analysis (Table 2). General criteria for gene selection included the chromosome locations (preferably evenly separated across the entire genome), functions of the encoded proteins (preferably conserved and well characterized), presence in all the strains as a single copy, and size of at least 1 kb (convenience of PCR primer design). In addition, *pgm* was selected based on the results of a previous study on *L. plantarum* (de las Rivas *et al.*, 2006), while *ftsZ* has been shown to be polymorphic in several LAB strains (Zhang & Dong, 2005). Selection of the remaining loci (*polA*, *mutL*, *metRS* and *nrpD*) was based on the presence of single nucleotide polymorphisms (SNPs) between *L. casei* ATCC 334 and *L. casei* 12A. These SNPs were identified in a previous study using comparative genome microarrays (H. Cai, J. R. Broadbent & J. L. Steele, unpublished data).

PCR amplification and DNA sequencing. Genomic DNA was extracted using an AquaPure Genomic DNA kit (Bio-Rad), with a 16–18 h proteinase K (final concentration, $100\text{ }\mu\text{g ml}^{-1}$; Invitrogen Life Technologies) treatment at 55°C , and it was stored at -20°C prior to use. PCR primers (Table 2) were designed using Primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi), on the basis of known gene sequences in *L. casei* ATCC 334. An approximately 800 bp internal fragment of each gene was amplified to allow accurate sequencing of a 600–700 bp fragment within each gene. PCR amplification was performed using iProof High-Fidelity DNA polymerase (Bio-Rad) with an iCycler Thermal Cycler (Bio-Rad). A single PCR programme was used for amplifications of all six housekeeping genes (initial denaturation at 98°C for 30 s, followed by 35 cycles of 98°C for 30 s, 60°C for 30 s, and 72°C for 30 s; final extension at 72°C for 10 min; and holding at 4°C). A 50 μl reaction was prepared according to iProof High-Fidelity DNA polymerase

Table 1. Origins and allelic profiles of the 40 *L. casei* strains analysed

Strain	ST	Allele						Origin	Reference or source*
		<i>pgm</i>	<i>polA</i>	<i>nrdD</i>	<i>metRS</i>	<i>mutL</i>	<i>ftsZ</i>		
L3	1	7	14	11	2	12	13	Human GI tract; USA	Walter <i>et al.</i> (2003)
L6	2	1	11	5	5	7	6	Human GI tract; USA	Walter <i>et al.</i> (2003)
L9	3	1	3	3	7	16	1	Human GI tract; USA	Walter <i>et al.</i> (2003)
L14	4	8	1	9	3	18	12	Human GI tract; USA	Walter <i>et al.</i> (2003)
L19	5	1	3	3	4	11	1	Human GI tract; USA	Walter <i>et al.</i> (2003)
L25	6	1	3	3	10	8	1	Human GI tract; USA	Walter <i>et al.</i> (2003)
L30	7	1	11	9	5	16	6	Human GI tract; USA	Walter <i>et al.</i> (2003)
CRF28	8	2	3	6	5	9	1	Human blood; USA	Accession no. AY299487
12A	1	7	14	11	2	12	13	Corn silage; WI, USA	J. L. Steele (unpublished)
32G	9	11	7	8	6	12	2	Corn silage; WI, USA	J. L. Steele (unpublished)
13/1	10	7	14	11	2	3	13	Corn silage; WI, USA	J. L. Steele (unpublished)
21/1	11	1	14	11	2	9	1	Corn silage; WI, USA	J. L. Steele (unpublished)
33/1	12	1	14	11	2	12	1	Corn silage; WI, USA	J. L. Steele (unpublished)
A2-309	13	3	6	10	5	6	4	Wine; Denmark	Rodas <i>et al.</i> (2005)
A2-362	14	9	8	4	5	5	16	Wine; Denmark	Rodas <i>et al.</i> (2005)
BI0231	15	1	3	3	5	2	1	Cucumber pickle; USA	USDA-ARS
USDA-P	16	8	5	7	5	3	5	Cucumber pickle; USA	USDA-ARS
UW-1	17	4	1	1	1	1	15	Cheese; WI, USA	J. L. Steele (unpublished)
UW-4	18	5	12	2	1	13	9	Cheese; WI, USA	M. Johnson
120501-1/6M	19	1	2	3	10	19	1	Cheese; WI, USA	M. Johnson
120501-3/6M	20	5	12	2	1	13	3	Cheese; WI, USA	M. Johnson
M36	21	1	3	3	10	19	1	Cheese; WI, USA	J. R. Broadbent (unpublished)
ATCC 334	22	12	6	10	5	4	7	Swiss-type cheese; USA	Chen <i>et al.</i> (2000)
ASCC 428	23	5	5	1	1	14	3	Cheddar cheese; Australia	I. Powell
ASCC 477	24	5	5	1	1	13	3	Cheddar cheese; Australia	I. Powell
ASCC 1087	25	5	10	2	1	13	11	Cheddar cheese; Australia	I. Powell
ASCC 1088	25	5	10	2	1	13	11	Cheddar cheese; Australia	I. Powell
ASCC 1123	26	5	10	2	1	13	3	Cheddar cheese; Australia	I. Powell
DPC 3971	27	10	1	8	3	10	12	Cheese; Ireland	Fitzsimons <i>et al.</i> (1999)
DPC 3968	26	5	10	2	1	13	3	Cheese; Ireland	Fitzsimons <i>et al.</i> (1999)
DPC 4108	28	1	3	3	5	15	1	Cheese; Ireland	Fitzsimons <i>et al.</i> (1999)
DPC 4249	26	5	10	2	1	13	3	Cheese; Ireland	Fitzsimons <i>et al.</i> (1999)
DPC 4748	29	6	10	2	1	13	3	Cheese; Ireland	Fitzsimons <i>et al.</i> (1999)
4R4	30	8	4	8	3	2	8	Cheese; Denmark	F. Vogensen
43M3	31	1	3	3	9	19	1	Cheese; Denmark	Adamberg <i>et al.</i> (2005)
7A1	32	5	5	1	1	20	14	Cheese; Denmark	Adamberg <i>et al.</i> (2005)
7R1	33	5	13	1	8	9	9	Cheese; Denmark	Christiansen <i>et al.</i> (2005)
83M4	34	5	5	1	1	17	10	Cheese; Denmark	Adamberg <i>et al.</i> (2005)
8I2	35	4	9	1	1	1	15	Cheese; Denmark	Adamberg <i>et al.</i> (2005)
MI280	36	10	1	8	3	3	12	Unknown	Unknown

*USDA-ARS, US Department of Agriculture, Agricultural Research Service. M. Johnson, Wisconsin Center for Dairy Research, Madison, WI, USA; I. Powell, Australian Starter Culture Research Center Limited, Werribee, Victoria 3030, Australia; F. Vogensen, Dept of Food Science, Royal Veterinary and Agricultural University, Frederiksberg C, Denmark.

directions. Following amplification, PCR mixtures were loaded on a 0.8% UltraClean agarose gel (Invitrogen Life Technologies), and separated by electrophoresis at 120 V for 1.5 h. The DNA bands (~800 bp) were excised from the gel, and purified using a Pure Link Quick Gel Extraction Kit (Invitrogen Life Technologies). DNA sequencing was performed with a BigDye Kit (Biotech Center, University of Wisconsin), using the following conditions: 35 cycles of 94 °C for 30 s, 50 °C for 20 s and 60 °C for 4 min; and holding at 4 °C. Sequencing products were purified with magnetic beads

(Beckman Coulter), and then sent to the Biotech Center for sequence determination.

MLST data analysis. Multiple sequence alignments were performed using molecular evolutionary genetic analysis (MEGA) software version 3.1 (<http://www.megasoftware.net>). Descriptive evolutionary analyses such as mol% G + C content, d_S/d_N ratios (where d_S is the number of synonymous substitutions per synonymous site, and d_N is the number of non-synonymous substitutions per non-synonymous site), and

Table 2. Genes and PCR primers

Gene	Gene function	PCR primer (5'→3')*	Size of amplicon (bp)
<i>ftsZ</i>	Cell division, Z-ring formation	GGCATTGCACAACCTGAAAGA; GCATCGTCTGCGTTAGTTTG	764
<i>polA</i>	DNA polymerase I	TTATCATGTGGCCGAACAAA; GTTTGCCTCAAAGTCTGCAA	858
<i>mutL</i>	DNA mismatch repair protein MutL	ATCGGCAACATTAAGCAACC; GATGACGCCCATTTGGATAAC	835
<i>metRS</i>	Methionyl-tRNA synthetase	CGGTATTTTGCCAGCCTTTA; CATTTCCGCTTTTAGCTTGC	742
<i>nrdD</i>	Anaerobic ribonucleoside-triphosphate reductase	GCTTGAAGCGTGATTAGCC; ACATTTCGATCGCCAATTGTT	815
<i>pgm</i>	Phosphoglucomutase	AGGCATTTGCTGCTCCTATG; GGGATCAGTCGCGATTAAGA	812

*Upper sequence, forward primer; lower sequence, reverse primer.

number of polymorphic sites and SNPs, were calculated using DnaSP version 4.0 (Rozas *et al.*, 2003). Different allelic sequences (with at least one nucleotide difference) were assigned arbitrary numbers. For each strain, the combination of six alleles defined its allelic profile, and a unique allelic profile was designated a sequence type (ST). The discrimination index (DI) value was calculated on the basis of numbers of allelic types (j), numbers of strains belonging to each type (n_j), and total numbers of strains analysed (N), as described by Hunter & Gaston (1988) with the following equation:

$$D = 1 - \frac{1}{N(N-1)} \sum_{j=1}^s n_j(n_j-1) \quad [1]$$

A minimum evolution (ME) tree for *L. casei* strains was constructed by using MEGA software version 3.1, based on the numbers of parsimoniously informative sites, and the results of a bootstrapping test of strain phylogeny (Kumar *et al.*, 2004). The numbers of synonymous substitutions per synonymous site were calculated from the concatenated nucleotide sequences using the modified Nei-Gojobori Jukes-Cantor method implemented in the MEGA program. The Reticulate program (Jakobsen & Easteal, 1996) was used to identify putative regions of recombination or gene conversion through the construction of a compatibility matrix. Split decomposition analysis was performed using the SplitsTree program (Huson, 1998).

RESULTS

API biochemical testing

Analysis of carbohydrate fermentation patterns by API biochemical testing demonstrated that all 40 *L. casei* strains could ferment galactose, glucose, fructose, mannose, mannitol, *N*-acetylglucosamine and tagatose, but they could not ferment glycerol, erythritol, arabinose, L-xylose, melibiose, raffinose, glycogen, xylitol, fucose, D-arabitol, potassium 2-ketogluconate and potassium 5-ketogluconate. Differences in carbohydrate utilization by *L. casei* strains are summarized in Table 3, and some niche-specific phenotypic traits were identified. For example, the ability to utilize some C₅ sugar alcohols (e.g. adonitol), C₅ sugars (e.g. ribose) and C₆ sugar

alcohols (e.g. sorbitol and dulcitol) was more prevalent in strains isolated from plant materials and human GI tracts than in cheese isolates. In contrast, the ability to ferment lactose was less common in strains isolated from plant materials than in those from cheese and human GI tracts.

Descriptive analysis of MLST loci and allelic diversity

Six widely distributed housekeeping gene loci (Fig. 1) were chosen from the core *L. casei* genome (approx. 2771 ORFs). A descriptive analysis of MLST for each locus is presented in Table 4. The MLST scheme revealed between 14 and 50 polymorphic sites in each gene, and a total of 199 SNPs in six loci. All six housekeeping-gene fragments had mol% G + C contents that were similar to the mean mol% G + C content of the *L. casei* genome (46.6%). The majority of SNPs in all six genes were synonymous. A premature stop codon was not found in any of the non-synonymous SNPs. The mean pairwise nucleotide difference per site (π /site), and the mean pairwise nucleotide difference per sequence (k), were calculated for each gene. The higher the π or k value, the higher the level of intragenic nucleotide polymorphism. The π /site values of the six genes varied from 0.00418 in *pgm* to 0.0276 in *metRS*. Similarly, *metRS* had the highest k value among the six loci (17.6).

Table 1 shows the allelic profiles and origins of all 40 *L. casei* strains analysed in this study. The number of alleles or allelic types per gene ranged from 10 (*metRS*) to 20 (*mutL*). Analysis of all six loci resulted in 36 STs, with a DI of 0.994. Generally, strains from the human GI tract, corn silage, wine and pickle displayed distinct allelic profiles at the six loci, except that L3 (a human GI tract strain) and 12A (a corn silage strain) shared identical alleles at all six loci. Two sets of cheese strains could not be differentiated by MLST. These included strains collected from Australia (ASCC 1087 and ASCC 1088), and strains collected from Australia (ASCC 1123) and Ireland (DPC 3968 and DPC 4249).

Table 3. Phenotypic differences in carbohydrate fermentation of *L. casei* strains

Substrate	Strains able to utilize substrate (%)			
	Cheese (n=22)	Plant (n=9)	GI (n=7)	Blood (n=1)
D-Ribose	77	100	100	100
D-Xylose	0	11	0	0
D-Adonitol	0	22	43	100
D-Sorbose	32	33	57	0
L-Rhamnose	0	11	0	0
Dulcitol	9	56	57	0
Inositol	0	0	14	0
D-Sorbitol	46	89	86	100
Methyl- α -D-manno- pyranoside	5	0	0	100
Methyl- α -D-gluco- pyranoside	9	22	29	0
Amygdalin	59	22	57	100
Arbutin	86	67	86	100
Salicin	96	100	100	100
D-Cellobiose	91	78	86	100
D-Maltose	100	78	100	100
D-Lactose	83	22	71	100
Sucrose	73	67	86	100
D-Trehalose	96	100	100	100
Inulin	27	67	86	100
D-Melezitose	96	78	100	100
Starch	0	11	0	0
Gentiobiose	64	78	57	100
D-Turanose	83	89	100	100
D-Lyxose	5	0	0	0
L-Arabitol	9	0	29	0
Potassium gluconate	14	11	0	0

Although *metRS* was determined to have the highest number of intragenic nucleotide polymorphisms, it was the least discriminatory gene for the 40 *L. casei* strains, as 23 of

the 40 *L. casei* strains shared identical alleles (either allele 1 or allele 5). The *metRS* allele 1 appeared to be specific to cheese-derived strains, whereas the *metRS* allele 5 was observed in strains from all ecological origins, other than cheese. In contrast, *mutL* was determined to have an intermediate level of intragenic nucleotide polymorphisms, but separated the 40 strains into the highest numbers of alleles ($n=20$). Therefore, *mutL* provided the highest discriminatory power for all 40 *L. casei* strains (DI 0.931), as well as for the 22 cheese-derived strains (DI 0.809).

Evidence for selection and recombination

Rates of synonymous and non-synonymous substitutions per site were estimated from concatenated allelic sequence alignments for each gene among the 40 *L. casei* strains (Table 3). The d_S/d_N ratio ranged from 33.6 for *nrdD* to 7.9 for *mutL*. Three genes (*polA*, *metRS* and *nrdD*) showed positive Tajima's D values (Tajima, 1989), indicating potential balancing selection in these genes, which was consistent with higher numbers of polymorphisms and d_S/d_N ratios.

To probe potential recombination, we used the Reticulate program (Jakobsen & Easteal, 1996), and constructed a compatibility matrix of 160 parsimoniously informative sites in the six gene fragments. Fig. 2 shows many highly incompatible sites between the six loci where nucleotide changes at these sites are inferred to have occurred multiple times, possibly due to recombination or repeated mutation (Jakobsen & Easteal, 1996). We used split decomposition analysis to detect possible conflicting phylogenetic signals (Bandelt & Dress, 1992). Evidence of recombination during evolution can also be detected when an interconnected network is displayed in the split graph (Huson, 1998). The split graphs of all six loci showed different network structures (Fig. 3a), suggesting intragenic recombination occurred during the evolution of these six loci. A combined split graph based on a distance matrix of pairwise distances of all alleles in the six loci also displayed a network-like structure, with several parallel paths indicative of the presence of incompatibilities resulting from recombination or recurrent mutation (Fig. 3b). Additionally, the combined split graph generated three major clusters that are consistent with the clusters in the MLST phylogeny tree (Fig. 4a). We have designated these groups clusters I, II and III, with cluster II representing most of the silage-derived strains, cluster III representing all cheese-derived strains, and cluster I representing the rest of the strains of various sources (Figs 3b and 4a).

MLST-based strain phylogeny, and estimation of evolutionary time scale

A consensus phylogeny using the ME algorithm based on the MLST data resolved three significant clusters with >70 % bootstrap support, and several other distinct

Fig. 1. Locations of 6 MLST loci in the *L. casei* ATCC 334 genome.

Table 4. Descriptive analysis of MLST data

Gene	Fragment analysed (bp)*	G+C content (mol%)	No. of			π /site†	k ‡	Tajima's <i>D</i> value	No. of		d_S/d_N §
			Polymorphic sites	SNPs	Alleles				Syn.	Nonsyn.	
<i>polA</i>	731 (27.0)	44.8	36	39	14	0.0147	10.7	0.590	31	8	13.1
<i>mutL</i>	747 (38.2)	49.8	29	29	20	0.00811	6.06	-0.381	21	8	7.9
<i>metRS</i>	636 (32.1)	47.5	50	50	10	0.0276	17.6	1.76	42	8	18.3
<i>nrdD</i>	735 (33.7)	48.7	44	45	11	0.0146	10.7	0.0545	41	4	33.6
<i>pgm</i>	734 (40.2)	46.9	14	14	12	0.00418	3.07	-0.213	11	3	11.8
<i>ftsZ</i>	676 (53.8)	49.3	26	26	16	0.00829	5.61	-0.281	22	4	18.8

*Percentage of the complete gene is given in parentheses.

†Mean pairwise nucleotide difference per site.

‡Mean pairwise nucleotide difference per sequence.

§Syn. synonymous sites; Nonsyn. non-synonymous sites.

branches among the SNP haplotypes (Fig. 4a). The deepest node in the ME phylogeny separated most of the cheese-derived strains from strains of the human GI tract and those of other food-related sources. The ME phylogeny provided consistent groupings with split decomposition (Fig. 3b).

To estimate the divergence time in different clusters of *L. casei*, we used the ME phylogeny for the 40 strains based on concatenated sequences of the six MLST loci (a combined total of 1419 allelic codons) that could be rooted with homologous genes in the closely related species *Pediococcus pentosaceus* (>90% nucleotide sequence identity over a minimum alignment length of 90% of both genes). Divergence times between different clusters are indicated by the scale of years in Fig. 4(a). Calculations were based on the number of single nucleotide substitutions in each

strain, and the estimated rate of single nucleotide substitutions between *Escherichia coli* and *Salmonella enterica* of 4.7×10^{-9} per site per year (Doolittle *et al.*, 1996; Lawrence & Ochman, 1998). Results indicated that the divergence of the three clusters of *L. casei* occurred approximately 1.5 million years ago, whereas most cheese and silage strains in clusters III and II seemed to have diversified more recently (Fig. 4a).

Comparison to PFGE

The 40 *L. casei* strains were analysed by PFGE, and a UPGMA tree was constructed based on *Sfi*I restriction patterns (Fig. 4b). PFGE discriminated all the strains, including those not differentiated by MLST. When compared with the ME tree, the PFGE tree showed a similar topology for the *L. casei* strains, including a relatively large cluster of cheese-derived strains. However, some human GI tract strains (L9 and L6) and wine strains (A2-309 and A2-362) seemed to be closely related to the main clusters of cheese strains on bifurcating branches in the PFGE tree, conflicting with relationships shown in the ME tree. Also similar to the ME tree, strains from blood, pickle, human GI tract and corn silage appeared to be genetically diverse, and grouped in different clusters. In both the ME tree and the PFGE tree, cheese strains did not cluster based on their geographical origin.

DISCUSSION

Lactobacillus species play a key role in the production of fermented foods and beverages. However, few studies have characterized strains of different ecological origins using both genotypic and phenotypic approaches. We have assembled and characterized a set of 40 *L. casei* strains that have different ecological and geographical origins. While an earlier comparison of complete genome sequences of nine *Lactobacillus* species revealed frequent

Fig. 2. Compatibility matrix of 160 parsimoniously informative SNPs in the six housekeeping genes. Highly incompatible sites are indicated by black squares.

Fig. 3. Split decomposition analysis of 40 *L. casei* strains based on concatenated sequences of six housekeeping genes. Formation of a parallelogram structure is suggestive of recombination. (a) Split decomposition of alleles for individual MLST loci. (b) Combined split decomposition of alleles for all six MLST loci.

gene loss and acquisitions, presumably via HGT (Makarova *et al.*, 2006), this study reports, for what we believe to be the first time, evidence that recombination and selective pressure are likely to have contributed to the evolution of *L. casei*, possibly facilitating adaptation to different ecological niches.

API biochemical testing identified some niche-specific carbohydrate-utilization patterns. For instance, lactose utilization is less prevalent in plant isolates than in those from cheese and human GI tracts, presumably due to relatively recent acquisitions of lactose metabolic genes, which are often plasmid encoded (Siezen *et al.*, 2005), in

Fig. 4. (a) Linearized ME tree based on 1419 allelic codons of the 40 *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 1000 random bootstrap replicates for the consensus tree. (b) UPGMA tree based on *Sfi*I-PFGE macrorestriction patterns. Geographical locations of cheese strains are labelled.

cheese-derived strains, and presumably in strains isolated from cheese- and milk-consuming human hosts, via HGT and subsequent natural selection.

PFGE provided higher discriminatory power than MLST on differentiation of *L. casei*

PFGE identifies large insertions, deletions and rearrangement of DNA, while MLST detects all the genetic variations within the amplified gene regions. Therefore, MLST is often found to provide better discriminatory ability than PFGE. However, in this study, although MLST provided good discriminatory power, differentiating 36 out of the 40 strains examined, PFGE was able to discriminate all the strains, including those that could not be separated by MLST. To improve the discriminatory power of MLST, we sequenced two additional genes (*gdh*, which encodes glutamate dehydrogenase, and *gyrB*, which encodes the β subunit of DNA gyrase) that have been reported to be polymorphic in a recent MLST study on *L. plantarum* (de las Rivas *et al.*, 2006); nevertheless, we could not separate the four strains not differentiated by the six-gene MLST analysis (data not shown). This suggests that portions of the *L. casei* genomes harbouring insertions, deletions and rearrangement have accumulated at higher rates than slowly evolving intragenic point mutations in the house-keeping genes. In fact, complete sequencing of *L. casei* ATCC 334 has revealed 130 complete or partial transposase genes, and two phage-related gene clusters (Makarova *et al.*, 2006; Ventura *et al.*, 2006). Also, LAB contain a relatively high number of plasmids, and the contribution of plasmid-encoded genes ranges from 0 to 4.8 % among the total gene contents in the fully sequenced LAB genomes (Makarova *et al.*, 2006). Furthermore, comparison of the complete genomes of multiple strains of different *Lactobacillus* species has also revealed extensive gene loss and acquisitions in *Lactobacillus* genomes, mainly via bacteriophage- and conjugation-mediated HGTs (Makarova *et al.*, 2006). Such genome events could be easily detected by PFGE, which is a DNA-banding-pattern-based method, but often they are missed by MLST.

Cluster analysis of *L. casei* suggests niche specificity

MLST data for six housekeeping genes allowed us to group *L. casei* strains into three clusters: a cheese cluster, a silage cluster, and a cluster with strains of different origins, but primarily those from human GI tracts and cheeses. Some correlation was observed when comparing the ME tree with the PFGE tree. The topological discrepancies between the ME tree and the PFGE tree could be explained by the fact that PFGE is more sensitive in detecting large insertions, deletions and genome rearrangements than MLST. Due to the unpredictable mutation rates of insertions or deletions in *L. casei* genomes, we interpreted genetic relatedness among *L. casei* strains solely based on the ME tree.

Compared with nucleotide sequence diversity of many Gram-positive food-borne pathogens, such as *Listeria monocytogenes* (Nightingale *et al.*, 2005), *L. casei* house-keeping genes are relatively conserved, reflected by lower π values in general. The mean rate of intragenic polymorphism of the MLST loci analysed in this study ranged from 1.4 % (*pgm*) to 7.8 % (*metRS*) among the 40 *L. casei* strains examined. This rate is even lower in cheese and silage strains, implying that *L. casei* strains isolated from the same ecological niche have less nucleotide sequence diversity, and are likely to have been exposed to similar selective pressures in that ecological niche. More interestingly, the low rate of nucleotide polymorphism appeared to be independent of the geographical locations from where these *L. casei* strains were isolated, as cheese isolates do not cluster based on their geographical origins, suggesting that environmental selective pressures for cheese strains are the same regardless of geographical origin.

***L. casei* has a recombinatorial population structure**

Even though *L. casei* strains are an industrially important LAB, with broad commercial applications (Mayra-Makinen & Bigret, 1998), their population structure has not been fully explored. Considerable reticulate evolution occurred between genes and network structures found in all six MLST loci by split decomposition, suggesting that many mutations are involved in parallel events, and that recombination in the MLST loci examined is frequent. These events may have facilitated rapid adaptation of *L. casei* to different environments. The existence of recombination is expected since many insertion sequences and several bacteriophage-associated genomic regions have been identified in the fully sequenced *L. casei* ATCC 334 genome (Makarova *et al.*, 2006; Ventura *et al.*, 2006), providing opportunities for exchange of genetic materials. This is also consistent with previous reports that other *Lactobacillus* species display a recombinatorial population structure. For example, strong evidence for intraspecies recombination was observed in *L. plantarum* by both presence of network structure in split decomposition analysis and linkage equilibrium (de las Rivas *et al.*, 2006).

Although a high degree of recombination, and a high level of phylogenetic heterogeneity among the 40 *L. casei* strains, were observed, cheese strains in cluster III in both the ME tree (Fig. 4a) and the combined split graph (Fig. 3b) seemed to be clonal. This suggests that the cheese-derived *L. casei* strains in cluster III may have a common recent ancestor, despite having been isolated from different geographical locations, probably because dairy farming in both the USA and Australia are linked to immigration from Europe (Denmark and Ireland), and thus the common ancestor of these strains has been carried to different cheese plants around the world, and become a stable contaminant in a specific cheese plant.

Selective pressure was detected in the *L. casei* housekeeping genes

The housekeeping genes examined by MLST had mol% G + C contents that were similar to that of the rest of the *L. casei* genome. This suggests that these genes have been present in *L. casei* for a long period of time, rather than being recently acquired through HGT.

A majority of synonymous mutations (d_s/d_N of >1) indicates the predominance of a purifying selection, preferentially associated with elimination of variations in amino acids. In this study, the high d_s/d_N ratio (33.6) observed for *nrdD* is suggestive of strong purifying selective pressure (selection against non-synonymous substitutions at the DNA level). This value is similar to those estimated by using whole genome sequences of *Lactobacillus gasseri* and *Lactobacillus johnsonii* (38.5 ± 0.5); these sequences reflected an unusually high mutation rate of the *Lactobacillus* species because of the intense evolutionary pressure (Makarova *et al.*, 2006). Synonymous and non-synonymous substitutions in housekeeping genes can arise from random nucleotide mutations or intragenic recombination events via HGT. In this study, the majority of SNPs were found to be synonymous. Some of the non-synonymous SNPs could possibly lead to adaptive niche expansion, and provide a selective advantage for *L. casei* to survive non-conventional habitats. However, a more in-depth functional characterization will be necessary to elucidate the potential effects of these non-synonymous substitutions on protein structure and functionality, and their correlation to bacterial adaptation to different environmental niches.

Additionally, Tajima's *D* tests detected positive values on genes *polA*, *metRS* and *nrdD*. A positive Tajima's *D* value is an indication of a history of positive Darwinian selection, most likely to balance selection (to maintain the genetic polymorphisms within a population) on protein-coding genes in bacterial genomes. These three genes were also found to have high levels of nucleotide polymorphisms. Surprisingly, however, they were also the least discriminatory (generated the fewest alleles) of the six genes examined. A plausible explanation for the contradiction between high sequence polymorphisms and low discriminatory power found in the allelic profiles of the 40 *L. casei* strains examined is that many strains shared identical nucleotide sequences or alleles in these genes. This suggests that either these genes tend to avoid substantial diversification, or missense mutations in these genes leading to attenuated functionality have been purged by natural selection during *L. casei* evolution.

Divergence of different genetic clusters of *L. casei* was relatively recent

Based on the 199 SNPs found in this study, we estimate that the major lineages of *L. casei* diverged approximately 1.5 million years ago. Compared with the speciation time

frame between *E. coli* and *Salmonella*, about 100 million years ago (Lawrence & Ochman, 1998), the diversification of these clusters within the *L. casei* species is relatively recent. In particular, divergence of cheese clusters seems very recent. This is consistent with the fact that cheese is a relatively new ecological niche, as cheese manufacture is believed to have begun approximately 8000 years ago (Fox & McSweeney, 2004). The recent intraspecies divergence of *L. casei* could have resulted from changes in its ecology, such as host shifts and adaptation to new environmental niches. Genome degradation (such as loss of ancestral genes) and metabolic simplification may have also contributed to the lineage diversification of *L. casei* populations (Makarova *et al.*, 2006). A more balanced strain selection for each ecological niche may increase the strength of the conclusions with respect to adaptive evolution towards specific niches. Further to this, more in-depth genomic and proteomic studies of additional *L. casei* strains should shed new insights on the evolution and geographical dissemination of this industrially important species.

ACKNOWLEDGEMENTS

We thank Ron Agee for technical assistance with PFGE, and Lenese Grant for help with DNA sequencing. We thank Finn Vogensen (Dept of Food Science, Royal Veterinary and Agricultural University, Frederiksberg C, Denmark), Mark Johnson (Wisconsin Center for Dairy Research, Madison, WI, USA), Tom Beresford (Teagasc, Oak Park Research Centre, Carlow, Ireland), Ian Powell (Australian Starter Culture Research Center Limited, Werribee, Victoria 3030, Australia), Kurt Reed (Marshfield Clinic Research Foundation, Marshfield, WI, USA), Gerald Tannock (Dept of Microbiology and Immunology, University of Otago, Dunedin, New Zealand) and Fred Breidt (Dept of Food Science, North Carolina State University, Raleigh, NC, USA) for providing the *L. casei* strains. Funding has been provided for this research and publication from Dairy Management, Inc. through the Center for Dairy Research, the College of Agricultural and Life Sciences at the University of Wisconsin, and the USDA Cooperative State Research, Education and Extension Service (CSREES) project WIS04908.

REFERENCES

- Adamberg, K., Antonsson, M., Vogensen, F. K., Nielsen, E. W., Kask, S., Moller, P. L. & Ardo, Y. (2005). Fermentation of carbohydrates from cheese sources by non-starter lactic acid bacteria isolated from semi-hard Danish cheese. *Int Dairy J* 15, 873–882.
- Bandelt, H. J. & Dress, A. W. M. (1992). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol Phylogenet Evol* 1, 242–252.
- Broadbent, J. R., Houck, K., Johnson, M. E. & Oberg, C. J. (2003). Influence of adjunct use and cheese microenvironment on nonstarter lactic acid bacteria populations in Cheddar-type cheese. *J Dairy Sci* 86, 2773–2782.
- Chen, H., Lim, C. K., Lee, Y. K. & Chan, Y. N. (2000). Comparative analysis of the genes encoding 23S–5S rRNA intergenic spacer regions of *Lactobacillus casei*-related strains. *Int J Syst Evol Microbiol* 50, 471–478.
- Christiansen, P., Petersen, M. H., Kask, S., Moller, P. L., Petersen, M., Nielsen, E. W., Vogensen, F. K. & Ardo, Y. (2005). Anticlostridial

- activity of *Lactobacillus* isolated from semi-hard cheeses. *Int Dairy J* 15, 901–909.
- de las Rivas, B., Marcobal, A. & Munoz, R. (2004). Allelic diversity and population structure in *Oenococcus oeni* as determined from sequence analysis of housekeeping genes. *Appl Environ Microbiol* 70, 7210–7219.
- de las Rivas, B., Marcobal, A. & Muñoz, R. (2006). Development of a multilocus sequence typing method for analysis of *Lactobacillus plantarum* strains. *Microbiology* 152, 85–93.
- Doolittle, R. F., Feng, D., 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.
- Enright, M. C. & Spratt, B. G. (1999). Multilocus sequence typing. *Trends Microbiol* 7, 482–487.
- Fitzsimons, N. A., Cogan, T. M., Condon, S. & Beresford, T. (1999). Phenotypic and genotypic characterization of non-starter lactic acid bacteria in mature cheddar cheese. *Appl Environ Microbiol* 65, 3418–3426.
- Fox, P. F. & McSweeney, P. L. H. (2004). Cheese: an overview. In *Cheese Chemistry, Physics and Microbiology*, pp. 1–37, vol. 1, 3rd edn. Edited by P. F. Fox, P. L. H. McSweeney, T. M. Cogan & T. P. Guinee. California: Elsevier.
- Germond, J. E., Lapierre, L., Delley, M., Mollet, B., Felis, G. E. & Dellaglio, F. (2003). Evolution of the bacterial species *Lactobacillus delbrueckii*: a partial genomic study with reflections on prokaryotic species concept. *Mol Biol Evol* 20, 93–104.
- Hunter, P. R. & Gaston, M. A. (1988). Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. *J Clin Microbiol* 26, 2465–2466.
- Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 14, 68–73.
- Jakobsen, I. B. & Easteal, S. (1996). A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Comput Appl Biosci* 12, 291–295.
- Kandler, O. & Weiss, N. (1986). Genus *Lactobacillus*. In *Bergey's Manual of Systematic Bacteriology*, vol. 2, 9th edn, pp. 1063–1065. Edited by P. H. A. Sneath, N. S. Mair, M. E. Sharpe & J. G. Holt. Baltimore: Williams & Wilkins.
- Kumar, S., Tamura, K. & Nei, M. (2004). MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* 5, 150–163.
- Lacher, D. W., Steinsland, H., Blank, T. E., Donnenberg, M. S. & Whittam, T. S. (2007). Molecular evolution of typical enteropathogenic *Escherichia coli*: clonal analysis by multilocus sequence typing and virulence gene allelic profiling. *J Bacteriol* 189, 342–350.
- Lawrence, J. G. & Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci U S A* 95, 9413–9417.
- Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K. & other authors (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* 95, 3140–3145.
- Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., Pavlov, A., Pavlova, N., Karamychev, V. & other authors (2006). Comparative genomics of lactic acid bacteria. *Proc Natl Acad Sci U S A* 103, 15611–15616.
- Mayra-Makinen, A. & Bigret, M. (1998). Industrial use and production of lactic acid bacteria. In *Lactic Acid Bacteria – Microbiology and Functional Aspects*, 2nd edn, pp. 73–102. Edited by S. Salminen & A. V. Wright. New York: Marcel Dekker.
- Michael, R. W., Rodolphe, B. & Philippe, H. (2006). *Methods for typing Lactobacillus species in food products, dietary supplements or animal feed by PCR amplification of CRISPR repeats*. PCT Int Appl 48 pp. CODEN: PIXXD2 WO 2006073445 A2 20060713 CAN 145:118272 AN 2006:681305 CAPLUS.
- Miragaia, M., Thomas, J. C., Couto, I., Enright, M. C. & de Lencastre, H. (2007). Inferring a population structure for *Staphylococcus epidermidis* from multilocus sequence typing (MLST) data. *J Bacteriol* 189, 2540–2552.
- Molenaar, D., Bringel, F., Schuren, F. H., de Vos, W. M., Siezen, R. J. & Kleerebezem, M. (2005). Exploring *Lactobacillus plantarum* genome diversity by using microarrays. *J Bacteriol* 187, 6119–6127.
- Nightingale, K. K., Windham, K. & Wiedmann, M. (2005). Evolution and molecular phylogeny of *Listeria monocytogenes* isolated from human and animal listeriosis cases and foods. *J Bacteriol* 187, 5537–5551.
- Olvera, A., Cerdà-Cuellar, M. & Aragon, V. (2006). Study of the population structure of *Haemophilus parasuis* by multilocus sequence typing. *Microbiology* 152, 3683–3690.
- Rodas, A. M., Ferrer, S. & Pardo, I. (2005). Polyphasic study of wine *Lactobacillus* strains: taxonomic implications. *Int J Syst Evol Microbiol* 55, 197–207.
- Rozas, J., Sánchez-DeLarrio, J. C., Messegue, X. & Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19, 2496–2497.
- Siezen, R. J., Renckens, B., van Swam, I., Peters, S., van Kranenburg, R., Kleerebezem, M. & de Vos, W. M. (2005). Complete sequences of four plasmids of *Lactococcus lactis* subsp. *cremoris* SK11 reveal extensive adaptation to the dairy environment. *Appl Environ Microbiol* 71, 8371–8382.
- Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595.
- Tynkkynen, S., Satokari, R., Saarela, M., Mattila-Sandholm, T. & Saxelin, M. (1999). Comparison of ribotyping, randomly amplified polymorphic DNA analysis, and pulsed-field gel electrophoresis in typing of *Lactobacillus rhamnosus* and *L. casei* strains. *Appl Environ Microbiol* 65, 3908–3914.
- Vasquez, A., Ahrne, S., Pettersson, B. & Molin, G. (2001). Temporal temperature gradient gel electrophoresis (TTGE) as a tool for identification of *Lactobacillus casei*, *Lactobacillus paracasei*, *Lactobacillus zeae* and *Lactobacillus rhamnosus*. *Lett Appl Microbiol* 32, 215–219.
- Ventura, M., Canchaya, C., Bernini, V., Altermann, E., Barrangou, R., McGrath, S., Claesson, M. J., Li, Y., Leahy, S. & other authors (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.
- Walter, J., Heng, N. C., Hammes, W. P., Loach, D. M., Tannock, G. W. & Hertel, C. (2003). Identification of *Lactobacillus reuteri* genes specifically induced in the mouse gastrointestinal tract. *Appl Environ Microbiol* 69, 2044–2051.
- Zhang, B. & Dong, X. Z. (2005). Partial sequence homology of FtsZ in phylogenetics analysis of lactic acid bacteria. *Wei Sheng Wu Xue Bao* 45, 661–664.

Edited by: T. Abee