Multilocus sequence typing of *Lactobacillus casei* isolates from naturally fermented foods in China and Mongolia

Qiuhua Bao, Yuqin Song, Haiyan Xu, Jie Yu, Wenyi Zhang, Bilige Menghe, Heping Zhang, and Zhihong Sun¹

Key Laboratory of Dairy Biotechnology and Engineering, Ministry of Education, School of Food Science and Engineering, Inner Mongolia Agricultural University, Huhhot 010018, China

ABSTRACT

Lactobacillus casei is a lactic acid bacterium used in manufacturing of many fermented food products. To investigate the genetic diversity and population biology of this food-related bacterium, 224 *Lb. casei* isolates and 5 reference isolates were examined by multilocus sequence typing (MLST). Among them, 224 *Lb. casei* isolates were isolated from homemade fermented foods, including naturally fermented dairy products, acidic gruel, and Sichuan pickles from 38 different regions in China and Mongolia. The MLST scheme was developed based on the analysis of 10 selected housekeeping genes (*carB*, *clpX*, *dnaA*, *groEL*, *murE*, *pyrG*, *pheS*, *recA*, *rpoC*, and *uvrC*). All 229 isolates could be allocated to 171 unique sequence types, including 25 clonal complexes and 71 singletons. The high index of association value (1.3524) and standardized index of association value (0.1503) indicate the formation of an underlying clonal population by all the isolates. However, split-decomposition, relative frequency of occurrence of recombination and mutation, and relative effect of recombination and mutation in the diversification values confirm that recombination may have occurred, and were more frequent than mutation during the evolution of *Lb. casei*. Results from Structure analyses (version 2.3; <http://pritch.bsd.uchicago.edu/structure.html>) demonstrated that there were 5 lineages in the *Lb. casei* isolates, and the overall relatedness built by minimum spanning tree showed no clear relationship between the clonal complexes with either the isolation sources or sampling locations of the isolates. Our newly developed MLST scheme of *Lb. casei* was an easy and valuable tool that, together with the construction of an MLST database, will contribute to further detailed studies on the evolution and population genetics of *Lb. casei* from various niches.

Key words: naturally fermented foods, *Lactobacillus casei*, multilocus sequence typing, genetic diversity

INTRODUCTION

Lactic acid bacteria (LAB) belong to a large family of gram-positive bacteria that can ferment glucose to lactic acid as the major metabolic end product. Lactic acid bacteria are often responsible for food preservation and flavor development, and some LAB are highly beneficial to human health. *Lactobacillus casei* is an important LAB species that is important in the production of many fermented foods. It is often present in traditional fermented food products, such as cheese, qula (Bao et al., 2012a,b), koumiss (Sun et al., 2010b), kurut (Sun et al., 2010a), pickle (Yu et al., 2012), and fermented acidic gruel (Yu et al., 2011a). Most importantly, these bacteria can enter the host's gastrointestinal tract, and thus effectively confer beneficial effects within the host gut; in particular, they can improve nutrition and aid disease prevention and therapy. Thus, some *Lb. casei* isolates are regarded as probiotics and are consumed specifically because of the potentially desirable properties they confer on the host (Zhang et al., 2014).

Accurate identification and an understanding of the genetic diversity and population biology of LAB isolates is essential both for basic research in bacteriology and for continued improvement of the food industry, particularly the dairy industry. In the past, LAB species have been classified using phenotypic identification methods, particularly the sugar fermentation test (Wang et al., 2008). However, phenotypic tests often give ambiguous results. Subsequently, many researchers have been trying to discriminate them using molecular methods, such as 16S rRNA gene sequencing (Bao et al., 2012a), species-specific PCR (Bao et al., 2012a), random fragment length polymorphism (Bao et al., 2012a), SDS-PAGE profiles (Gatti et al., 2001), random amplified polymorphic DNA-PCR, and amplified fragment length polymorphism (Di Cagno et al., 2010). Although these methods have advantages, their reliability in distinguishing between similar isolates or

Received January 5, 2016.

Accepted March 28, 2016.

¹Corresponding author: sunzhihong78@163.com

species is still limited, and this hinders studies on the population biology of LAB.

With the development of gene sequencing technology and a decrease in the costs associated with sequencing, multilocus sequence typing (MLST) is increasingly regarded as the bacterial identification and typing method of choice. Multilocus sequence typing is principally based on the partial nucleotide sequences of multiple housekeeping genes (Tanigawa and Watanabe, 2011), is highly discriminatory, and provides unambiguous results because it is based directly on nucleotide sequences. This method was first described in 1998 (Maiden et al., 1998) and several publicly accessible MLST databases are now available, including MLST (<http://www.mlst.net/>) and PubMLST (<http://pubmlst.org/>).

The MLST method has been used in several important studies on genetic diversity and bacterial pathogen population biology (Tanabe et al., 2007; Usein et al., 2014). Several MLST schemes have also been developed for species of LAB, including *Lb. casei* (Cai et al., 2007; Diancourt et al., 2007), *Lactobacillus acidophilus* (Ramachandran et al., 2013), *Lactobacillus delbrueckii* (Tanigawa and Watanabe, 2011), *Lactobacillus plantarum* (de Las Rivas et al., 2006; Tanganurat et al., 2009), *Lactobacillus sanfranciscensis* (Picozzi et al., 2010), *Lactobacillus helveticus* (Sun et al., 2015), *Lactobacillus sakei* (Chaillou et al., 2013), *Lactococcus lactis* (Xu et al., 2014), and *Leuconostoc lactis* (Dan et al., 2014). In these studies, strains were classified with high resolution by using MLST based on the analysis of housekeeping gene sequences, including *aroE*, *ddl*, *dnaE*, *fusA*, *ftsZ*, *glnA*, *gyrB*, *glpF*, *gltX*, *gyrB*, *gpd*, *gdh*, *hemN*, *hsp60*, *ileS*, *lepA*, *leuS*, *ldhL*, *mutS*, *mutL*, *metRS*, *nrdD*, *pepV*, *pgm*, *polA*, *recA*, *recG*, *recP*, *xpt*, *yqil*, *tkt*, and *tpi*. These published reports clearly demonstrated that MLST analysis is a powerful tool for genotyping and for population biology studies of specific bacterial species.

With respect to *Lb. casei*, this species has high genotypic diversity, numerous subspecies and its population characteristics within the natural environment are unclear (Dobson et al., 2004). For this reason, we aimed to develop more effective MLST scheme for *Lb. casei* using 10 housekeeping genes on isolates from a wide range of naturally fermented foods from China and Mongolia. Such a scheme should facilitate interesting insights into the genetic diversity, population biology, and evolutionary relationships among the isolates evaluated.

MATERIALS AND METHODS

Bacterial Isolates

A total of 224 *Lb. casei* isolates were selected from the LAB Culture Collection of the Inner Mongolia

Agriculture University. They had been identified using various methods including conventional phenotypic identification, 16S rRNA sequence, and species-specific PCR (Airidengcaিকে et al., 2010; Yu et al., 2011a,b, 2012; Bao et al., 2012a,b; Liu et al., 2012). These included *Lb. casei* isolates from several naturally fermented dairy foods, including yogurt (fermented cow milk), kurut (fermented yak milk), koumiss (fermented mare milk), whey, cream, qula (a kind of traditional cheese), and other naturally fermented nondairy foods, such as Sichuan pickles and acidic gruel (a popular traditional fermented food in the western regions of in Inner Mongolia in China). The products from which the isolates came originated from 38 different regions in 5 provinces (Inner Mongolia, Xinjiang, Tibet, Sichuan, and Gansu) in China, and from 2 cities (Bulgan and Zavkhan) in Mongolia between 2002 and 2009. In addition, sequences of 5 reference isolates (ATCC 334, Zhang, 12A, DSM5622, and N1115) from the NCBI genome database (<http://www.ncbi.nlm.nih.gov/genome/652>) were also used. The whole genome of *Lb. casei* Zhang was sequenced in our laboratory (Zhang et al., 2010). Information for all isolates is listed in Supplemental Table S1 (<http://dx.doi.org/10.3168/jds.2016-10857>).

These *Lb. casei* isolates were freeze-dried in a cryoprotectant (10% skim milk broth containing 0.1% sodium glutamate) before being stored at -80°C before use. As a large proportion of the studied isolates (133) were from Inner Mongolia, which is a vast territory covering 700,000 km² in northern China, the region was subdivided into 3 zones. The first zone was Inner Mongolia A, which covered the west of the region including the cities of Bayinnaoer, Ordos, and Baotou. The second zone was Inner Mongolia B, which covered the central part of the region including the cities of Hohhot and Xilingol. The final zone was Inner Mongolia C, which covered the east of the region including the cities of Hulunbeier and Chifeng.

DNA Extraction

All *Lb. casei* isolates were retrieved from storage and each was grown in 5 mL of de Man, Rogosa, Sharpe broth (Difco, Becton Dickinson and Company, Franklin Lakes, NJ) at 37°C for 18 to 22 h without shaking. Bacteria from each isolate were harvested by centrifugation and resuspended in 500 μL of TE solution (pH 8.0, 10 mM Tris hydrochloride, 1 mM EDTA). Each suspension was put through 3 cycles of freezing and thawing in liquid nitrogen and a 65°C water bath. Genomic DNA was extracted by a modified cetyltrimethylammonium bromide (CTAB) method as described previously (Bao et al., 2012a). Purified DNA was diluted to a final concentration of 100 ng/ μL for further evaluation.

Selection of MLST Loci

Ten housekeeping loci (*carB*, *clpX*, *dnaA*, *groEL*, *murE*, *pyrG*, *pheS*, *recA*, *rpoC*, and *uvrC*) were selected for MLST analysis of the *Lb. casei* isolates. General criteria for selection of housekeeping genes include their location on the chromosome, the function of the encoded proteins, their presence in all isolates as a single copy, and that gene size is at least 1 kb (Cai et al., 2007). The 10 housekeeping genes encode the following proteins: chromosomal replication initiator protein *dnaA* (*dnaA*), CTP synthase (*pyrG*), RNA polymerase β subunit (*rpoC*), recombinase A (*recA*), ATP-dependent protease ATP-binding subunit ClpX (*clpX*), carbamoyl phosphate synthetase (*carB*), UDP-N-acetylmuramyl tripeptide synthase (*murE*), phenylalanyl-tRNA synthetase subunit α (*pheS*), chaperonin GroEL (*groEL*), and excinuclease ABC subunit C (*uvrC*). The primers for amplifying these 10 genes were designed by Primer Premier 5.0 (Premier Biosoft, Palo Alto, CA) using the published *Lb. casei* Zhang genome sequences (Zhang et al., 2010). The information of genes and primers used for MLST of *Lb. casei* are listed in Table 1.

PCR Amplification and DNA Sequencing

Each selected gene locus was amplified using PCR with the corresponding primers (Table 1) using the chromosomal DNA of *Lb. casei* isolates as a template. The PCR reactions were done in final volumes of 50 μ L in an automatic thermal cycler (PTC-200, MJ Research,

Waltham, MA). Thermal cycling conditions were an initial denaturation at 94°C for 3 min.; 30 cycles of 94°C for 1 min, annealing at 52 to 60°C for 1 min and 72°C for 30 s (annealing temperature for each locus is listed in Table 1), and a final elongation step at 72°C for 10 min. The PCR products were analyzed by electrophoresis in 1.2% agarose gels. Amplicons of the correct size were purified with a DNA Purification Kit (TransGen Biotech Company, Beijing, China). Both DNA strands of the extracted gene fragments were sequenced with an ABI 3730XL genetic analyzer (service provided by Shanghai Sangni Biosciences Corporation of China, Shanghai, China). The same primers were used in both the PCR and DNA sequencing reactions.

MLST Data Analysis

For each gene, the ClustalW package of software MEGA 6.0 (version 6.0, www.megasoftware.net; Tamura et al., 2013) was used to align all the sequences. Then the sequences were trimmed to a uniform length by manually deleting the jagged bases for further comparison and analyses. Definition of the alleles was achieved using a nonredundant data set and minimum spanning tree (MSTree) analysis in the BioNumerics Software (version 6.6, Applied-Maths, Sint Maartens-Latem, Belgium), with default settings. Allelic sequences with more than one nucleotide difference were given different arbitrary numbers. For each of the 10 MLST loci or alleles, a unique nucleotide sequence was assigned. The

Table 1. Genes and primers used for multilocus sequence typing of *Lactobacillus casei*

Locus	Gene product	PCR primer	Sequence (5'-3')	Location (bp)	Amplicon length (bp)	Annealing temperature (°C)
<i>carB</i>	Carbamoyl phosphate synthetase	<i>carB</i> _F <i>carB</i> _R	ATTTGCCTTGATCGTAA TGCTACTTTCATAGGTGCTGT	1409893 to 1409115	779	52
<i>clpX</i>	ATP-dependent protease ATP-binding subunit ClpX	<i>clpX</i> _F <i>clpX</i> _R	TTTGCTCGAAATTACGCT CGAATAAAGAAGACCTGGATA	1296258 to 1296827	570	55
<i>dnaA</i>	Chromosomal replication initiator protein <i>dnaA</i>	<i>dnaA</i> _F <i>dnaA</i> _R	TCTGTTTATTTATGGTGGCG CTGCGGTCATCAAGTTTCA	438 to 1052	615	54
<i>groEL</i>	Chaperonin GroEL	<i>groEL</i> _F <i>groEL</i> _R	GTTGCGTCCGTTTCTTCCT GATCCGATACTTGCGTTCC	2175965 to 2175219	747	60
<i>murE</i>	UDP-N-acetylmuramyl tripeptide synthase	<i>murE</i> _F <i>murE</i> _R	ATGATTCACGCAAGGTCTC AGCACTGGCAAACAGGTAG	236665 to 237374	710	52
<i>pyrG</i>	CTP synthase	<i>pyrG</i> _F <i>pyrG</i> _R	CAGTGTAAAGGAATTGCGTAG CAGGTTGCGTCTCAGTTGTA	2481603 to 2480960	615	54
<i>pheS</i>	Phenylalanyl-tRNA synthetase subunit α	<i>pheS</i> _F <i>pheS</i> _R	GTGATCGAAGCCCGCAAAG CGGCACCTAACACTTCAATCCA	1627618 to 1626964	655	58
<i>recA</i>	Recombinase A	<i>recA</i> _F <i>recA</i> _R	GGCTGGAATGCTGATG GTTGATGTATTTGCGGTTGA	848247 to 848921	675	58
<i>rpoC</i>	RNA polymerase β subunit	<i>rpoC</i> _F <i>rpoC</i> _R	TTGTCGGCTATCAACTGCT TCAACCTTGGTGCCCTTTT	2429852 to 2429102	751	60
<i>uvrC</i>	Excinuclease ABC subunit C	<i>uvrC</i> _F <i>uvrC</i> _R	CAGTTAGAACGCCACAGC CTGCCGATGGAAGGTAAT	1297004 to 1297626	623	60

sequence types (ST) were defined by the unique allelic profiles (consisting of the allele numbers at each of the 10 MLST loci). The same ST was applied to several isolates when they shared the same allelic profiles. A group of ST in which 2 connected ST differed by 1 locus were defined as clonal complexes (CC) using the eBURST program (Feil et al., 2004) within the BioNumerics program.

Descriptive evolutionary analyses, such as mol% G+C content, d_N/d_S ratios (where d_N is the number of nonsynonymous substitutions per nonsynonymous site and d_S is the number of synonymous substitutions per synonymous site), nucleotide diversity (π) per site, the number of polymorphic sites, index of association (I_A) and standardized index of association (I_A^S) value, were calculated using the Start version 2.0 program (Jolley et al., 2001). The split decomposition method was used to assess the degree of tree-like structure for alleles of each locus and all ST using SplitsTree 4.0 (Huson and Bryant, 2006). The phylogenetic trees were constructed using the maximum likelihood statistical method in the MEGA 6.0 software package, as described previously.

The software Structure v2.3 (<http://pritchardlab.stanford.edu/software.html>) with an admixture model (Falush et al., 2007) was used to identify the ancestral subpopulations and assign ancestry proportions for each ST. The K value that generated the highest posterior probability was used as the probable number of ancestral populations. Three individual runs per value of K (chosen between 3 and 15) were performed using 200,000 burn-in iterations and 300,000 sampling iterations. The ρ/θ (relative frequency of occurrence of recombination and mutation) and r/m (relative effect of recombination and mutation in the diversification) was accounted for using ClonalFrame v1.2 (Didelot and Falush, 2007).

Nucleotide Sequence Accession Numbers

Allele sequences of the 10 MLST loci were deposited in GenBank under accession numbers KT800420-KT802659.

RESULTS

Gene Sequencing

We developed a new MLST scheme for *Lb. casei* based on fragments of 10 housekeeping genes. Ten widely distributed housekeeping gene loci (Supplemental Figure S1; <http://dx.doi.org/10.3168/jds.2016-10857>) were chosen from the core *Lb. casei* Zhang genome (approximately 2,804 open reading frames in genome 28.6184 Mb). The DNA fragments of the 10 housekeeping genes were successfully amplified for each of the 224 *Lb. casei* isolates. The gene sequences obtained for each *Lb. casei* isolate were trimmed, aligned, and analyzed. Nucleotide positions of the sequences showing ambiguities were excluded from the analysis. Ultimately, fragment sizes of the 10 housekeeping genes ranged from 422 (*clpX*) to 603 bp (*carB*) (Table 2) and were used for MLST analysis.

Allelic Variation in 10 Housekeeping Genes

Nucleotide and allelic variation of the 10 genes were successfully analyzed by MLST software for all isolates selected in this study (Table 2). The number of alleles per locus ranged from 16 to 31. Partial sequencing of 10 loci revealed 237 SNP across the 4,920 bp of the concatenated sequences. The number of polymorphic sites per locus ranged from 13 (*pyrG*) to 36 (*carB*). The G+C content of the 10 housekeeping genes ranged from

Table 2. Nucleotide and allelic diversity of the multilocus sequence typing loci

Locus	Application Size (bp)	Number of		G + C content (%)	π (Nucleotide diversity)	d_N/d_S ¹
		Alleles	Polymorphic sites			
<i>carB</i>	603 ²	31	36	48.07	0.00725	0.1147
<i>clpX</i>	422	22	20	44.89	0.0076	0.0600
<i>dnaA</i>	456	19	15	41.78	0.00819	0.0036
<i>groEL</i>	483	24	23	46.16	0.00662	0.0442
<i>murE</i>	508	20	19	46.19	0.00614	0.1040
<i>pheS</i>	475	23	26	44.58	0.00882	0.0507
<i>pyrG</i>	474	16	13	46.58	0.00585	0.0626
<i>recA</i>	484	27	47	47.69	0.00813	0.4896
<i>rpoC</i>	554	20	14	47.83	0.00567	0.0459
<i>uvrC</i>	461	21	24	48.13	0.01211	0.0666
Concatenated	4920	171	237	46.19	0.00479	—

¹ d_N = number of nonsynonymous changes per nonsynonymous site; d_S = number of synonymous changes per synonymous site.

²The maximum and minimum values in each column are shown in bold.

41.78 (*dnaA*) to 48.13% (*uvrC*), the mean mol% G+C content (46.19%) of the 10 housekeeping gene fragments was similar to the mol% G+C content of the *Lb. casei* Zhang genome (46.5%; Zhang et al., 2010). The average nucleotide diversity per site (π) was 0.0048% for all the concatenated sequences taken together and ranged from 0.0057 (*rpoC*) to 0.0121 (*uvrC*). The d_N/d_S ratio ranged from 0.0036 for *dnaA* to 0.4896 for *recA*.

Evidence for Recombination

Linkage disequilibrium refers to nonrandom association of alleles at different loci and it was expressed as I_A and I_A^S . In our study an I_A of 1.3524 ($P = 0.000$) and I_A^S of 0.1503 ($P = 0.000$) was obtained for the 10 loci, which indicated a tendency for linkage disequilibrium between the alleles at the level of the whole population. This also indicated there was a clonal structure among *Lb. casei* strains in our study.

Split decomposition analysis was done to assess the degree of tree-like structure present based on the allelic found for each locus. Split graphs showed that all loci, except for the *carB* gene, had a slight parallelogram structure (Supplemental Figure S2; <http://dx.doi.org/10.3168/jds.2016-10857>). It is generally accepted that genes with a parallelogram structure had undergone little intergenic recombination during evolution. In addition, the concatenated sequences of the 10 loci displayed a typical extensive network-like structure (Supplemental Figure S2), suggesting several recombination events based on the phi test ($P = 4.266E-11$). We deduced that 9 genes with the most recombination history might have contributed most to the recombination events. The high value for r/m ($r/m = 1.549$, 95% credibility interval = 0.7144–2.4026) and ρ/θ ($\rho/\theta = 1.54$, 95% credibility interval = 0.6336–2.5566), indicated that mutation occurred less frequently than recombination at these loci during the evolution of *Lb. casei*.

Assignment of ST in *Lb. casei*

The 229 *Lb. casei* isolates (224 plus 5 reference isolates) were genotyped by combining (4,920 bp) the 10 gene loci in a MLST scheme and could be divided into 171 ST (Supplemental Table S1; <http://dx.doi.org/10.3168/jds.2016-10857>). According to source, the 224 isolates could be divided into 3 groups. The first group contained 172 strains isolates from traditional fermented dairy foods from China and Mongolia and were assigned to 140 ST (ST4~ST111 and ST143~ST171). The second group contained 45 *Lb. casei* isolates from acidic gruel from the western area of Inner Mongolia,

China, and were assigned to 25 ST (ST112~ST136). The final group was represented by 7 *Lb. casei* isolates from fermented pickles from Sichuan province and were assigned to 6 ST (ST137~ST142). In addition, 5 genome strains were identified as a single ST, the exceptions being N1115 and Zhang. Strain 12A belonged to ST1, ATCC334 belonged to ST2, DSM 5622 belonged to ST3, Zhang belonged to ST4, and strain N1115 belonged to ST6.

Among the 171 ST, the most represented ST were ST85, composed of 8 isolates (3.5% of all isolates) from fermented yak milk products of the Gansu, Sichuan, and Tibet provinces. Six ST were represented by at least 4 isolates: ST6 (4 isolates), ST39 (4 isolates), ST112 (5 isolates), ST118 (4 isolates), ST120 (4 isolates), and ST121 (5 isolates). A total of 18 ST contained 2 to 3 *Lb. casei* isolates. In the remaining 143 ST (83.6% of all ST), each ST corresponded to single *Lb. casei* isolate.

Isolate Relationships Based on Allelic Profiles

The relatedness of the different ST was investigated using the program eBURST. Clonal complexes were defined as groups of profiles differing by no more than one gene from at least one other profile of the group (Rahman et al., 2014). Overall, a total of 171 ST were clustered into 25 CC and 71 singletons [Figure 1 and Supplemental Figure S3 (<http://dx.doi.org/10.3168/jds.2016-10857>)]. The 138 isolates member of the 25 CC accounted for 60.3% of all the *Lb. casei* isolates studied. The colored zones between some circle groups indicate profiles belonging to the same CC (Figure 1). Clonal complex 1 was represented by the most ST (17), including ST78, ST98, ST160, ST148, ST97, ST143, ST159, ST80, ST102, ST157, ST85, ST86, ST165, ST167, ST166, ST81, and ST168. Sequence type 78 was predicted to be the founder genotype of CC1 and contained 2 isolates [IMAU80728 from Gansu and IMAU20567 from Mongolia; Supplemental Table S1 (<http://dx.doi.org/10.3168/jds.2016-10857>)]. Clonal complex 2 contained the second largest number of ST (9), of which ST20 was considered to be the founder genotype. Similarly, CC3 to CC9 each contained of at least 4 ST; ST6, ST115, ST19, ST49, ST29, ST145 and ST 110 were considered as the founder genotypes of CC3 to CC9, respectively (Supplemental Figure S3). Moreover, reference strains DSM 5622, Zhang, and N1115 were located in CC12, CC2, and CC3, respectively.

To explore the microgenetic evolutionary relationships among the 229 *Lb. casei* isolates from different regions and sources, an MSTree was obtained based on allelic profiles at the concatenated sequence of the

Figure 1. Minimum spanning tree analysis of the 224 *Lactobacillus casei* isolates from different regions and 5 reference isolates based on allelic profiles of the 10 housekeeping genes. Each circle corresponds to a sequence type (ST), and the circle size denotes the number of isolates sharing the same ST. Colored zones between some groups of circles indicate that these profiles belong to the same clonal complex (CC). The strength of the link (bold, plain, or discontinuous) is related to the genetic similarity (number of common alleles) between profiles.

10 gene loci using the software program BioNumerics 6.6 (Figure 1 and Figure 2). Each circle corresponds to a ST, and the size of the circle is related to the number of isolates found within that specific ST. Circles were colored based on the different geographic region or source foodstuff from which the *Lb. casei* were isolated. The MSTree based on region (Figure 1) showed

that the 229 *Lb. casei* isolates could be divided into 3 major branches. Branch A mainly included isolates from Tibet (blue circles) and Inner Mongolia zone C (green circles) in China, and reference strains ATCC 334, N1115, and DSM 5622. Branch B mainly included isolates from Inner Mongolia zone A (red circles) and C (green circles) and reference strains Zhang and 12A.

Figure 2. Minimum spanning tree analysis of the 224 *Lactobacillus casei* isolates from different fermented foods and 5 reference isolates based on allelic profiles of the 10 housekeeping genes.

In branch C, some isolates from Sichuan, Gansu, and Mongolia were closely related to some isolates from the Tibetan region; however, they seemed to be distant from isolates from Inner Mongolia zone C. In addition, some isolates from Inner Mongolia zone C were also closely related to those from the Tibetan region (see branch A). In summary, most isolates from Inner Mongolia zone A (red circles) belonged to branch B, whereas most isolates from Tibet (blue circles) were

distributed in branches A and C. Most isolates from Gansu (yellow circles) were distributed in branches A and C. All isolates from Mongolia were found in branch C.

To analyze the relationship between the CC and the source dairy products for the 229 *Lb. casei* isolates, an MSTree based on source was constructed (Figure 2). Most of the isolates from koumiss and acidic gruel were distributed in branches A and B (Figure 2), whereas all

isolates from qula were distributed in branches C and A. Two isolates from cream (2/2 isolates), 7 from pickle (7/7 isolates), and 8 from Kurut (8/9 isolates) were found in branches A, B, and C, respectively. Isolates from yogurt and whey were also found in branches A, B, and C. In conclusion, we noted no significant associations between branches and regions or the type of dairy products from which the collection of *Lb. casei* isolates originated.

Population Structure and Phylogenetic Relationships

The Structure linkage model was applied to the sequence data sets of 171 ST. A maximal posterior probability of $K = 5$ was found in our sample, meaning that these *Lb. casei* ST were divided into 5 ancestral lineages (Figure 3A). Lineage 1 (green) contained 20 ST, mainly from yogurt and koumiss from the eastern area of Inner Mongolia. Lineage 2 (blue) and lineage 3 (red) comprised 33 and 46 ST, respectively, and most of these isolates were from fermented milk products from Mongolia, Gansu, Sichuan, Tibet, and Inner Mongolia. Lineage 4 (yellow) consisted of 22 ST, mainly from pickle and yogurt from Inner Mongolia. Lineage 5 (orange) consisted of 50 ST derived from central Mongolia, Gansu, and Tibet. Furthermore, the 5 reference isolates were found distributed within 3 lineages: *Lb. casei* Zhang and *Lb. casei* 12A were located in lineage 1; *Lb. casei* N1115 and *Lb. casei* DSM 5622 were located in lineage 5; and *Lb. casei* ATCC 334 showed great homology with isolates in lineage 3 (Figure 1). We observed little admixture of ancestral sources between these 5 lineages, suggesting high homogeneity of the ST in each lineage.

The neighbor-joining (NJ) tree was constructed based on the 171 ST determined by concatenating the sequences from the 10 loci (Figure 3B). The colored branches of the NJ tree are in accordance with the color of the groups identified using Structure. The NJ tree revealed 7 branches (Figure 3B). Lineages 1 to 4 were consistent with the 5 subpopulations identified from the population structure analysis; however, we found that the ST in lineage 5 (orange) were distributed in differentiated branches of the NJ tree and was divided into 3 small lineages 5A to C. Most branches within all the 5 lineages were poorly supported by bootstrap values, indicating inherently incongruent phylogenetic signals in the NJ topology. Therefore, we deduced that the 229 *Lb. casei* isolates formed 5 lineages during evolution based on the combined evidence from the Structure analyses and NJ tree.

DISCUSSION

Multilocus sequence typing is increasingly used as a routine typing tool that enables international comparison of prokaryotic isolates. It has not only been applied to epidemiological studies of many bacterial pathogens, but also in the studies of population structure and genetic evolution of other bacteria (Cai et al., 2007; Rahman et al., 2014; Dan et al., 2015; Sun et al., 2015). In our study, a new MLST scheme for *Lb. casei* was developed based on 10 gene fragments to characterize the genetic population diversity and evolutionary relationships among 224 *Lb. casei* isolates from homemade fermented foods in China and Mongolia.

In this MLST scheme, 10 gene targets (*carB*, *clpX*, *dnaA*, *groEL*, *murE*, *pheS*, *pyrG*, *recA*, *uvrC*, and *rpoB*) were selected based on previous MLST studies of LAB and the *Lb. casei* Zhang genome. These genes have been successfully used in the molecular typing of *Lb. fermentum* (Dan et al., 2015), *Lb. helveticus* (Sun et al., 2015), *Lc. lactis* (Xu et al., 2014), *Le. lactis* (Dan et al., 2014), and *Le. mesenteroides* (Zhang et al., 2015). From these 10 gene loci, the 224 *Lb. casei* isolates and 5 reference isolates could be allocated to 171 ST. The number of alleles per locus ranged from 16 to 31. When any gene locus with the lowest number of alleles was removed from the analysis, the number of ST decreased. Both Cai et al. (2007) and Diancourt et al. (2007) reported MLST schemes for *Lb. casei* to analyze the nucleotide variability at 6 to 7 genetic loci of 40 to 52 *Lb. casei* isolates from human and plant material and dairy foods. Compared with these previous MLST schemes, we identified an average of 17 distinguishable alleles per locus among the 171 ST identified, which is more than the average of 8 to 14 alleles observed from 40 and 52 *Lb. casei* isolates. The existence of such diversity can be due in part to the fact that we selected more loci and more isolates form a wide range of geographic and dairy origin. Moreover, these results demonstrate that our MLST scheme using 10 genes provides better discriminatory power.

The d_N/d_S ratio was calculated as a measure of the degree of selection in the population and as a measure of selective pressure at the protein level. In our analysis of the divergences in the 10 housekeeping genes, the d_N/d_S ratios (0.0036~0.4896) were much lower than 1, indicating a strong purifying selection which did not affect the AA composition. This is a typical phenomenon for housekeeping genes and is desired in MLST schemes. The nucleotide diversity (π value) ranged from 0.00567 in *rpoC* to 0.01211 in *uvrC*. The π values found in our study were higher than that of *Lb. casei*, ranging from 0.0051 to 0.0096 in a previously published

Figure 3. Population structures and neighbor-joining tree of 171 *Lactobacillus casei* sequence types (ST). (A) Sources of ancestry of each unique ST from 5 ancestral subpopulations by program Structure v2.3 analysis (<http://pritch.bsd.uchicago.edu/structure.html>; assuming $K = 5$ populations and applying the linkage model). Five ancestral lineages are colored in green (lineage 1), blue (lineage 2), red (lineage 3), yellow (lineage 4), and orange (lineage 5). Each ST is represented by a single line with the ST designation at the top consisting of colored stacked bars that indicate the proportion of ancestry from each of 6 subpopulations. (B) Neighbor-joining tree constructed from concatenated sequences of the 10 MLST loci. Bootstrap values are indicated for all branches; ST are colored according to their affiliation to 1 of the 5 ancestral lineages.

report (Diancourt et al., 2007), indicating a higher degree of intragenic nucleotide polymorphism in our selected housekeeping genes. Recombination is widely accepted that plays a driving role in the evolution of most bacteria (Didelot and Maiden, 2010). For test the recombination in the evolution of *Lb. casei*, I_A , I_A^S , r/m , and ρ/θ values and split decomposition analysis were used in our study. The high I_A value (1.3524) and I_A^S value (0.1503) identified an underlying clonal population structure for all the isolates. However, the network structure of the split-graph in our study suggested that recombination might have occurred; the high ρ/θ and r/m value also indicated that recombination occurred more frequently than mutation at these loci during the evolution of *Lb. casei*. Recombination may have facilitated rapid adaptation of *Lb. casei* to different environments, but it was not frequent enough to disrupt the clonal population structure. These results were consistent with previous reports for *Lb. casei* (Cai et al., 2007; Diancourt et al., 2007).

Population structure analysis using Structure suggested that the 5 ancestral lineages for our isolates contained a degree of admixture (Figure 3A). However, the NJ tree (Figure 3B) revealed that lineage 5 from Structure was further divided into 3 small lineages, although these were not well supported because of the low bootstrap values. This fact implied that the NJ method was unsuitable for analyzing MLST data from our *Lb. casei* isolates. The consistently low bootstrap values in NJ trees has also been reported in *Lb. sakei* (Chaillou et al., 2013). For these reasons, it has been suggested that approaches like eBURST and MS, which are less sensitive to the effects of recombination, would be more useful (Diancourt et al., 2007).

In this study, MSTree and eBURST analyses were used to explore the microgenetic evolutionary relationships among the 229 *Lb. casei* isolates from different regions and sources. The results found no significant associations between lineages and geographic locality or source within our collection of *Lb. casei* isolates (Figures 1 and 2). Some isolates from the same location and sources did cluster together to a certain extent. Reference strain Zhang, in particular, was isolated from central Inner Mongolia, so it belongs to ST4 along with the 2 Inner Mongolia and 1 Xinjiang isolates. However, a great number of isolates were dispersed across branches. Strains 12A and ATCC 334 from the United States were identified as a single ST and are very different from each other. Our findings are consistent with the results obtained previously for related species used in food fermentation processes such as *Lc. lactis*, *Lb. sanfranciscensis*, and *Lb. plantarum*, where no signifi-

cant clustering associated with the geographic origin of the isolates was noted (Tanganurat et al., 2009; Passerini et al., 2010; Picozzi et al., 2010). However, Cai et al. (2007) did report that a collection of 40 *Lb. casei* isolates examined were grouped into a cheese cluster, a silage cluster, and a mixed cluster, the latter primarily consisting of isolates from the human gastrointestinal tract and cheese isolates. Some correlation was also observed between the MS tree and the pulsed-field gel electrophoresis tree, suggesting high specificity for the individual ecological niches. We inferred that the discrepancy between our data and this literature may be related to the loci we selected and the *Lb. casei* isolates used; in our study only 2 to 9 isolates in total were isolated from cream, qula, whey, and pickle, whereas 123 and 45 isolates were from yogurt and acidic gruel, respectively. As more and more *Lb. casei* isolates are collected from different regions and sources, further evidence may become available for global grouping of isolates, which would provide strong indications of the factors that have affected the evolutionary history of this species.

CONCLUSIONS

In the present study, we developed a MLST scheme for *Lb. casei* and used it for typing analysis and for investigating the population biology of a set of 224 isolates. Using a large number of isolates allowed us to better interpret the ecological differences underlying the clonal structures observed. However, the isolates employed in our study were not correlated with geographic or food-source origin. Analysis of a geographically diverse and representative collection of isolates using MLST could provide a better understanding of the evolution of the *Lb. casei* genome and provide information for future studies on the structure and genetic evolution of *Lb. casei* globally.

ACKNOWLEDGMENTS

This research was supported by the Scientific Research Projects of Hi-Tech Research and Development Program of China (Beijing, China; 2011AA100902), the National Natural Science Foundation of China (Beijing, China; 31301516), Inner Mongolia Doctoral Scientific Research Foundation (Huhhot, China; BJ2013D-18), International S&T Cooperation Program of China (Beijing, China; ISTCP, 2014DFR31150), and Institution of Higher Education in Inner Mongolia (Huhhot, China; NJZY12096).

REFERENCES

- Airidengcaিকে, X. Chen, X. Du, W. Wang, J. Zhang, Z. Sun, W. Liu, L. I. Li, T. Sun, and H. Zhang. 2010. Isolation and identification of cultivable lactic acid bacteria in traditional fermented milk of Tibet in China. *Int. J. Dairy Technol.* 63:437–444.
- Bao, Q., J. Yu, W. Liu, M. Qing, W. Wang, X. Chen, F. Wang, M. Li, H. Wang, Q. Lv, and H. Zhang. 2012a. Predominant lactic acid bacteria in traditional fermented yak milk products in the Sichuan Province of China. *Dairy Sci. Technol.* 92:309–319.
- Bao, Q., W. J. Liu, J. Yu, W. H. Wang, M. J. Qing, X. Chen, F. Wang, J. C. Zhang, W. Zhang, J. M. Qiao, T. S. Sun, and H. P. Zhang. 2012b. Isolation and identification of cultivable lactic acid bacteria in traditional yak milk products of Gansu Province in China. *J. Gen. Appl. Microbiol.* 58:95–105.
- Cai, H., B. T. Rodriguez, W. Zhang, J. R. Broadbent, and J. L. Steele. 2007. Genotypic and phenotypic characterization of *Lactobacillus casei* strains isolated from different ecological niches suggests frequent recombination and niche specificity. *Microbiology* 153:2655–2665.
- Chaillou, S., I. Lucquin, A. Najjari, M. Zagorec, and M. C. Champomier-Verges. 2013. Population genetics of *Lactobacillus sakei* reveals three lineages with distinct evolutionary histories. *PLoS ONE* 8:e73253.
- Dan, T., W. Liu, Y. Song, H. Xu, B. Menghe, H. Zhang, and Z. Sun. 2015. The evolution and population structure of *Lactobacillus fermentum* from different naturally fermented products as determined by multilocus sequence typing (MLST). *BMC Microbiol.* 15:107.
- Dan, T., W. Liu, Z. Sun, Q. Lv, H. Xu, Y. Song, and H. Zhang. 2014. A novel multi-locus sequence typing (MLST) protocol for *Leuconostoc lactis* isolates from traditional dairy products in China and Mongolia. *BMC Microbiol.* 14:150.
- de Las Rivas, B., A. Marcobal, and R. Munoz. 2006. Development of a multilocus sequence typing method for analysis of *Lactobacillus plantarum* strains. *Microbiology* 152:85–93.
- Di Cagno, R., G. Minervini, E. Sgarbi, C. Lazzi, V. Bernini, E. Neviani, and M. Gobbetti. 2010. Comparison of phenotypic (Biolog System) and genotypic (random amplified polymorphic DNA-polymerase chain reaction, RAPD-PCR, and amplified fragment length polymorphism, AFLP) methods for typing *Lactobacillus plantarum* isolates from raw vegetables and fruits. *Int. J. Food Microbiol.* 143:246–253.
- Diancourt, L., V. Passet, C. Chervaux, P. Garault, T. Smokvina, and S. Brisse. 2007. Multilocus sequence typing of *Lactobacillus casei* reveals a clonal population structure with low levels of homologous recombination. *Appl. Environ. Microbiol.* 73:6601–6611.
- Didelot, X., and D. Falush. 2007. Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175:1251–1266.
- Didelot, X., and M. C. Maiden. 2010. Impact of recombination on bacterial evolution. *Trends Microbiol.* 18:315–322.
- Dobson, C. M., B. Chaban, H. Deneer, and B. Ziola. 2004. *Lactobacillus casei*, *Lactobacillus rhamnosus*, and *Lactobacillus zeae* isolates identified by sequence signature and immunoblot phenotype. *Can. J. Microbiol.* 50:482–488.
- Falush, D., M. Stephens, and J. K. Pritchard. 2007. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. *Mol. Ecol. Notes* 7:574–578.
- Feil, E. J., B. C. Li, D. M. Aanensen, W. P. Hanage, and B. G. Spratt. 2004. eBURST: Inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J. Bacteriol.* 186:1518–1530.
- Gatti, M., M. E. Fornasari, and E. Neviani. 2001. Differentiation of *Lactobacillus delbrueckii* ssp. *bulgaricus* and *Lactobacillus delbrueckii* ssp. *lactis* by SDS-PAGE of cell-wall proteins. *Lett. Appl. Microbiol.* 32:352–356.
- Huson, D. H., and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23:254–267.
- Jolley, K. A., E. J. Feil, M. S. Chan, and M. C. Maiden. 2001. Sequence type analysis and recombinational tests (START). *Bioinformatics* 17:1230–1231.
- Liu, W., Q. Bao, Jirimutu, M. Qing, Siriguleng, X. Chen, T. Sun, M. Li, J. Zhang, J. Yu, M. Bilige, T. Sun, and H. Zhang. 2012. Isolation and identification of lactic acid bacteria from Tarag in Eastern Inner Mongolia of China by 16S rRNA sequences and DGGE analysis. *Microbiol. Res.* 167:110–115.
- Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt. 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. USA* 95:3140–3145.
- Passerini, D., C. Beltramo, M. Coddeville, Y. Quentin, P. Ritzenthaler, M. L. Daveran-Mingot, and P. L. Bourgeois. 2010. Genes but not genomes reveal bacterial domestication of *Lactococcus lactis*. *PLoS ONE* 5:e15306.
- Picozzi, C., G. Bonacina, I. Vigentini, and R. Foschino. 2010. Genetic diversity in Italian *Lactobacillus sanfranciscensis* strains assessed by multilocus sequence typing and pulsed-field gel electrophoresis analyses. *Microbiology* 156:2035–2045.
- Rahman, A., C. Cailliez-Grimal, C. Bontemps, S. Payot, S. Chaillou, A. M. Revol-Junelles, and F. Borges. 2014. High genetic diversity among strains of the unindustrialized lactic acid bacterium *Carnobacterium maltaromaticum* in dairy products as revealed by multilocus sequence typing. *Appl. Environ. Microbiol.* 80:3920–3929.
- Ramachandran, P., D. W. Lacher, E. A. Pfeiler, and C. A. Elkins. 2013. Development of a tiered multilocus sequence typing scheme for members of the *Lactobacillus acidophilus* complex. *Appl. Environ. Microbiol.* 79:7220–7228.
- Sun, Z., W. Liu, W. Gao, M. Yang, J. Zhang, L. Wu, J. Wang, B. Menghe, T. Sun, and H. Zhang. 2010a. Identification and characterization of the dominant lactic acid bacteria from kurut: The naturally fermented yak milk in Qinghai, China. *J. Gen. Appl. Microbiol.* 56:1–10.
- Sun, Z., W. Liu, Y. Song, H. Xu, J. Yu, M. Bilige, H. Zhang, and Y. Chen. 2015. Population structure of *Lactobacillus helveticus* isolates from naturally fermented dairy products based on multilocus sequence typing. *J. Dairy Sci.* 98:2962–2972.
- Sun, Z., W. Liu, J. Zhang, J. Yu, W. Zhang, C. Cai, B. Menghe, T. Sun, and H. Zhang. 2010b. Identification and characterization of the dominant lactobacilli isolated from koumiss in China. *J. Gen. Appl. Microbiol.* 56:257–265.
- Tamura, K., G. Stecher, D. Peterson, A. Filipski, and S. Kumar. 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30:2725–2729.
- Tanabe, Y., F. Kasai, and M. M. Watanabe. 2007. Multilocus sequence typing (MLST) reveals high genetic diversity and clonal population structure of the toxic cyanobacterium *Microcystis aeruginosa*. *Microbiology* 153:3695–3703.
- Tanganurat, W., B. Quinquis, V. Leelawatcharamas, and A. Bolotin. 2009. Genotypic and phenotypic characterization of *Lactobacillus plantarum* strains isolated from Thai fermented fruits and vegetables. *J. Basic Microbiol.* 49:377–385.
- Tanigawa, K., and K. Watanabe. 2011. Multilocus sequence typing reveals a novel subspecies of *Lactobacillus delbrueckii*. *Microbiology* 157:727–738.
- Usein, C. R., M. Militaru, V. Cristea, and M. Straut. 2014. Genetic diversity and antimicrobial resistance in *Streptococcus agalactiae* strains recovered from female carriers in the Bucharest area. *Mem. Inst. Oswaldo Cruz* 109:189–196.
- Wang, J., X. Chen, W. Liu, M. Yang, Airidengcaিকে, and H. Zhang. 2008. Identification of *Lactobacillus* from koumiss by conventional and molecular methods. *Eur. Food Res. Technol.* 227:1555–1561.
- Xu, H., Z. Sun, W. Liu, J. Yu, Y. Song, Q. Lv, J. Zhang, Y. Shao, B. Menghe, and H. Zhang. 2014. Multilocus sequence typing of *Lactococcus lactis* from naturally fermented milk foods in ethnic minority areas of China. *J. Dairy Sci.* 97:2633–2645.
- Yu, J., X. Du, W. Wang, J. Zhang, W. Liu, Z. Sun, T. Sun, and H. Zhang. 2011a. Phenotypic and genotypic characteristics of lactic acid bacteria isolated from sour congee in Inner Mongolia of China. *J. Gen. Appl. Microbiol.* 57:197–206.

- Yu, J., W. Gao, M. Qing, Z. Sun, W. Wang, W. Liu, L. Pan, T. Sun, H. Wang, N. Bai, and H. Zhang. 2012. Identification and characterization of lactic acid bacteria isolated from traditional pickles in Sichuan, China. *J. Gen. Appl. Microbiol.* 58:163–172.
- Yu, J., W. H. Wang, B. L. Menghe, M. T. Jiri, H. M. Wang, W. J. Liu, Q. H. Bao, Q. Lu, J. C. Zhang, F. Wang, H. Y. Xu, T. S. Sun, and H. P. Zhang. 2011b. Diversity of lactic acid bacteria associated with traditional fermented dairy products in Mongolia. *J. Dairy Sci.* 94:3229–3241.
- Zhang, J., L. Wang, Z. Guo, Z. Sun, Q. Gesudu, L. Kwok, Menghebilige, and H. Zhang. 2014. 454 pyrosequencing reveals changes in the faecal microbiota of adults consuming *Lactobacillus casei* Zhang. *FEMS Microbiol. Ecol.* 88:612–622.
- Zhang, W., W. Liu, Y. Song, H. Xu, B. Menghe, H. Zhang, and Z. Sun. 2015. Multilocus sequence typing of a dairy-associated *Leuconostoc mesenteroides* population reveals clonal structure with intragenic homologous recombination. *J. Dairy Sci.* 98:2284–2293.
- Zhang, W., D. Yu, Z. Sun, R. Wu, X. Chen, W. Chen, H. Meng, S. Hu, and H. Zhang. 2010. Complete genome sequence of *Lactobacillus casei* Zhang, a new probiotic strain isolated from traditional homemade koumiss in Inner Mongolia, China. *J. Bacteriol.* 192:5268–5269.