International Food Research Journal 27(5): 883 - 892 (October 2020)
Journal homepage: *http://www.ifrj.upm.edu.my*

Comparative genomic analysis and proteolytic properties of *Lactobacillus curieae* CCTCC M2011381

[1]Chen, Y., [1,4]*Xie, J., [1]Sun, J., [2]Zhou, Z., [1]Zhang, R., [1,3]Li, H. and [1,4]Wei, D.

[1]*State Key Laboratory of Bioreactor Engineering, Department of Food Science and Engineering, East China University of Science and Technology, Shanghai 200237, P.R. China*
[2]*State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, P.R. China*
[3]*Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, P.R. China*
[4]*Shanghai Collaborative Innovation Centre for Biomanufacturing (SCICB), Shanghai 200237, P.R. China*

Article history

Received: 22 February 2020
Received in revised form: 6 June 2020
Accepted: 16 July 2020

Keywords

lactic acid bacteria, Lactobacillus curieae, comparative genomic analysis, proteolytic system

Abstract

Lactobacillus curieae CCTCC M2011381, a novel strain initially isolated from the brine of stinky tofu, acts as a starter in soymilk but not in milk. We investigated how evolution and physiology help *L. curieae* adapt to the soy environment. To do so, its whole genome was obtained through a large-scale sequencing using Illumina HiSeq 4000 and PacBio RSII. A comparative genome analysis between the strain and other four non-starter dairy strains, namely *Lactobacillus brevis* ATCC 367, *Pediococcus pentosaceus* ATCC 25745, *L. plantarum* WCFS1, and *L. reuteri* JCM 1112 was carried out based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Cluster of Orthologous Groups (COG) of protein databases. The results indicated that the assembled genome comprised one circular chromosome (2.1 Mb) without a plasmid. The strain was not only phylogenetically close to *L. brevis* ATCC 367, but also had a close COG distribution to *L. brevis* ATCC 367. The strain lacked cell envelope proteinase (CEP) which is used for the initiation of casein utilisation, and some oligopeptide transporter systems for the di-, tri-, and tetrapeptides (Dpp system) as detected by rapid annotation using the Subsystem Technology annotation sever. However, it did have an oligopeptide transporter (Opp) system, a peptidolysis system, and an amino acid synthetic pathway. The present work provides insight into the evolution of the *L. curieae* M2011381 proteolytic system, and its application in the fermentation of soy material.

Introduction

Lactic acid bacteria (LAB) are Gram-positive bacteria that are crucial players in majority of food fermentation ecosystems (Bourdichon *et al.,* 2012). It has already been confirmed that some LAB are probiotics that are beneficial for the host's health (Yang *et al.,* 2019; Ren *et al.,* 2020). Furthermore, fermentation with probiotics is one of the potential ways of improving the nutritional content of some foods (Melini *et al.,* 2019). For example, soy powder milk fermented with *Lactobacillus plantarum* P1201 contains more conjugated linoleic acid and isoflavone aglycones when compared with unfermented material (Xie et al., 2017). It is also known that *L. plantarum, L. sakei,* and *L. coryniformis* strains isolated from traditional Japanese fermented sushi and pickles could convert daidzin to daidzein when used as starters in the fermentation of soymilk (Tsuda and Shibata, 2017).

LAB inhabit a variety of ecological niches such as meat products, fruits, fermented and dairy products, as well as the gastrointestinal tract of humans and animals (Klaenhammer *et al.,* 2002; 2005; Horvath *et al.,* 2009) due to their ability to metabolise carbohydrates and proteins. The biochemical and genetic properties of the LAB proteolytic system in bovine milk have been extensively studied in recent years (Savijoki *et al.,* 2006; Griffiths and Tellez, 2013). Generally, this proteolytic system comprises three major components. First, the cell wall-bound proteinases initiate the degradation of extracellular casein into oligopeptides. Second, transporters take up the peptides into cells. Finally, intracellular peptidases degrade the peptides into shorter peptides and amino acids (Savijoki *et al.,* 2006; Liu *et al.,* 2010; De Angelis *et al.,* 2016). However, very little is known about the function of LAB in soybean. Some lactobacilli strains have reportedly been able to hydrolyse α- and α′-subunits of β-conglycinin;

*Corresponding author.
Email: jlxie@ecust.edu.cn

moreover, the strains show different proteolytic specificity in soy proteins (Aguirre *et al.,* 2008; 2014). Thus, a comparative genomic profile can provide evidence of the diversity of the proteolytic system in various LAB strains. For instance, the *L. acidophilus* group, including *L. acidophilus, L. johnsonii, L. gasseri, L. delbrueckii* subsp. *bulgaricus,* and *L. helveticus* strains, has been described as proteolytic system owners. The corresponding genes in these strains that encode for the transporters and enzymes have offered a wealth of information to further understand a novel LAB, the properties, and metabolic functions of its genome (Ramachandran *et al.,* 2013).

In the present work, the complete genome sequence of a novel LAB strain, *L. curieae* CCTCC M2011381 is presented, which was initially isolated from stinky tofu brine (Lei *et al.,* 2013), and could be used as starter for plant foods (Liu *et al.,* 2020). A comparative genome analysis was performed which included the profiles of evolution and physiological properties among *L. curieae* M2011381 and the other four non-starter dairy strains, which had a high genome similarity to the strain. Then, genetic information about the proteolytic system and the amino acid synthetic system of the strain were surveyed to study the specific characteristics of *L. curieae* M2011381 and its adaptation in soy niche.

Materials and methods

Strain and medium

Lactobacillus curieae CCTCC M2011381^T (M2011381) was originally isolated from stinky tofu brine and deposited in China Centre for Type Culture Collection (CCTCC, collection number M2011381). The strain was stored at -80°C in MRS medium supplemented with 20% (v/v) glycerol. The strain was aerobically cultured in MRS medium at 37°C under static conditions, or on MRS plates supplemented with 1.5% agar. Soy milk with 2.5% (w/w) protein content was purchased from Qingmei Green Food Co. Ltd. (Shanghai, China). Reconstituted skim milk which contained 3.0% protein was made by adding 10 g of the powder (Canpac Int. Ltd., New Zealand) into 100 mL water. All the cultures were maintained at 37°C. The cell count was determined by plating the culture sampled in different fermentation periods on the MRS plates. The pH of the culture was also recorded.

SDS-PAGE

The supernatants of milk and soymilk in different fermentation periods were obtained by centrifugation at 4°C, 8,000 *g* for 10 min. Protein degradation was analysed by SDS-PAGE using 8% stacking gel and 12.5% separating gels according to the PAGE Gel Fast Preparation Kit (EpiZyme Biotechnology Co. Ltd., Shanghai, China). Gel electrophoresis was alternately performed at 80 and 120 V. Gels were stained with 0.1% (w/v) Coomassie Brilliant Blue (R-250), and de-stained with a de-staining buffer (acetic acid:methanol:water = 2:1:17, v/v). An image of the gel was captured by Tanon 1600/1600R Gel Imaging System (Tanon Science and Technology Co. Ltd., Shanghai, China).

Genome sequencing and annotation

The whole genomic DNA was extracted from the strain M2011381 using Wizard Genomic DNA Purification Kit (Promega, Biotech Co., Ltd., Beijing, China) according to the manufacturer's protocol. The harvested DNA was detected by agarose gel electrophoresis and quantified by Qubit (Thermo Fisher Scientific, Waltham, MA). The whole genome sequence was obtained through large-scale sequencing using Illumina HiSeq 4000 and PacBio RSII by Beijing Genomics Institute (BGI, Shenzhen, China). The clean data was assembled by SOAP *de novo* software. GLIMMER 3.02 was applied to predict the gene sequence. Genes were functionally classified by COGs (https://www.ncbi.nlm.nih.gov/COG/). Other genome sequences used for comparative genome analysis were obtained from the NCBI microbial genome database, including *L. brevis* ATCC 367 (Accession No. CP000416), *P. pentosaceus* ATCC 25745 (Accession No. CP000422), *L. plantarum* WCFS1 (Accession No. AL935263), and *L. reuteri* JCM 1112 (Accession No. AP007281). All these genomes were rapidly annotated using the subsystem technology (RAST) annotation sever (http://rast.nmpdr.org/) for automated analysis.

Genome comparison and analysis

The KEGG (http://www.kegg.jp/) database and RAST annotation server were used for genome comparison (Kanehisa *et al.,* 2006; Aziz *et al.,* 2008). The proteinases were obtained from the non-redundant protein database UniProt (http://www.uniprot.org/) (Bairoch *et al.,* 2005); and they have been confirmed as members of the proteolytic system by experiments. A whole genome BLAST search against microbial proteins (E-value less than $1e^{-6}$, minimal alignment length percentage larger than 40%) was performed. The functional proteinases were found using the BlastP program (Altschul *et al.,* 1990). The phylogenetic analysis was conducted in MEGA5 with neighbour-joining (NJ) method (Tamura *et al.,* 2011). The genomic context was visualised using the Circos software (Krzywinski *et al.,* 2009). PHASTER

(PHAge Search Tool Enhanced Release, http://phaster.ca) was used for the rapid identification and annotation of prophage sequences within the genome (Zhou *et al.,* 2011; Arndt *et al.,* 2016). Orthologous genes were analysed by orthoMCL. The coding sequence (CDS) Venn diagram and COG frequency heat map with double hierarchical clustering were generated using RStudio, and the packages "pheatmap" and "VennDiagram". COG analysis was based on the phylogenetic classification of proteins encoded in the complete genomes (NCBI, www.ncbi.nlm.nih.gov-/COG/) project. To obtain singletons among the five strains, the whole protein sequence database was analysed by BlastP using the entire database, and was mapped and annotated with the eggNOG database. Statistically variable functions of different proteins were analysed by Fisher's exact test in BlastP (Jensen *et al.*, 2008).

Data availability

The complete genome sequence of *L. curieae* CCTCC M2011381 was deposited at the DDBJ/EMBL/GenBank database as sequencing project PRJNA266911, Accession No. CP018906.

Results and discussion

Growth properties of the strain M2011381

The growth properties of the strain M2011381 were monitored in three media: soymilk, MRS broth, and milk by testing the changes in viable cell numbers and pH during fermentation (Figure 1a). The strain could vigorously grow in soymilk. The cell count increased from log 7.1 to 12.7 CFU/mL during 24 h. Simultaneously, the pH of soymilk significantly decreased from 6.7 to 4.5. However, both the curves of cell count and pH in bovine milk were placid. The cell count changed from log 7.1 to 7.9 CFU/mL, and the pH declined from 6.3 to 5.9. MRS medium is suitable for most LAB; hence, the strain grew and reached the highest cell density of log 11.5 CFU/mL at 12 h in the medium. After that, the growth became stationary, and began to drop slightly. The pH of the MRS broth rapidly dropped from 6.5 to 4.6 during the first 12 h of the cell growth. However, the pH descent became slight at 24 h. Soymilk was thus concluded to be the optimal medium for strain growth. The composition of soymilk is different from milk. The utilisation of these two substrates needs different metabolic pathways and enzymes, which indicated that the genetic and metabolic properties of the strain M2011381 was more adapted to soy material than to milk. Moreover, there was higher cell growth in soymilk than in the other two substances, which indicated that this strain was expected to be the starter for soy food fermentation.

The proteolysis of soymilk and milk protein by the strain M2011381 during fermentation is displayed by SDS-PAGE (Figure 1b and 1c). Figure 1b shows that most soymilk proteins above 45 kDa were hydrolysed during the first 4 h of fermentation. The bands near 35 and 20 kDa are the acidic and basic subunits of glycinin, respectively, which are darker in the lane of unfermented soy milk. After the first 4 h, the bands of both subunits became extremely light, indicating that the glycinin was also hydrolysed by the strain. Peptides below 16 kDa accumulated at 4 h, which might be the hydrolysates of proteins in soymilk. However, milk proteins hardly changed during fermentation (Figure 1c). The results implied that the strain M2011381 possesses a strong proteolytic system in soy rather than milk protein.

Genome profile of the strain M2011381

The complete genome of *L. curieae* M2011381 was sequenced and annotated. The profile of the complete genome is shown in Table 1. The assembled genome comprised of one circular chromosome (2,095,860 bp) without any plasmid. This genome represents a typical size of lactobacilli (Makarova *et al.,* 2006). The average GC content of the complete genome was 39.8%. The complete genome contained 1,973 predicted coding sequences, 15 rRNA operons, and 62 tRNA-coding genes. The draft genome assembled into 29 contigs was previously sequenced and reported (Wang *et al.,* 2015). Most of the basic information about the genome is the same. The draft genome with a larger size also displayed more genes. However, the complete genome enveloped more predicted coding sequences, rRNA operons, and tRNA-coding genes when compared with the draft, which indicates that some of the non-repetitive sequences between the two genomes may hide in the gaps of the draft genome. PHASTER identified three regions that encoded for prophages in the genome of *L. curieae* M2011381, which were predicted by the program to be "incomplete" regions. Phage 1 (1,599,352 – 1,608,570 bp) contains 11 CDSs; phage 2 (1,676,604 – 1,685,668 bp) contains 7 CDSs; and phage 3 (1,829,246 – 1,860,778 bp) contains 14 CDSs.

Comparative genomics of the five strains

Using the complete genome of the strain M2011381 as a template, genomes of 24 LAB strains were chosen from GenBank by RAST analysis. Although the strains *L. brevis* subsp. *gravesensis* ATCC 27305, *L. hilgardii* ATCC 8290, and

Table 1. General genome features of *L. curieae* M2011381 and selected species.

	***L. curieae* CCTCC M2011381 (draft)**	***L. curieae* CCTCC M2011381 (complete)**	***L. brevis* ATCC 367**	***L. plantarum* WCFS1**	***L. reuteri* JCM 1112**	***P. pentosaceus* ATCC 25745**
Size (Mb)	2.19	2.10	2.29	3.31	2.04	1.83
No. of plasmids	0	0	2	3	0	0
GC content (%)	39.6	39.8	46.2	44.5	38.9	37.4
No. of proteins	1957	1973	2141	3013	1932	1710
No. of genes	2143	1994	2259	3124	2050	1797
tRNA genes	56	62	63	70	65	55
rRNA operons	6	15	15	15	18	15
Contigs	29	1	3	4	1	1

L. buchneri ATCC 11577 exhibited higher RAST scores, their genomes have not been assembled yet. Therefore, *L. brevis* ATCC 367, *P. pentosaceus* ATCC 25745, *L. plantarum* WCFS1, and *L. reuteri* JCM 1112 were selected for the comparison due to their well-assembled genomes. Moreover, these strains are non-starters in dairy (Kleerebezem *et al.,* 2003; Diep *et al.,* 2006; Morita *et al.,* 2008; Guo *et al.,* 2017). The general features of these four model bacteria are summarised in Table 1. The 16S rDNA sequences from 20 different lactobacilli strains were selected to construct a phylogenetic tree. The strain M2011381 was close to *L. brevis* ATCC 367, P. pentosaceus ATCC 25745, and L. plantarum WCFS1, and far from L. reuteri JCM 1112 in the phylogenetic tree. Differences of the RAST score were similar to the ranks of the phylogenetic evolution tree of the four strains, indicating that the selected model strains were suitable for further comparative analysis.

Among the five genomes, *P. pentosaceus* ATCC 25745 had the smallest genome and the fewest protein-coding genes. Some reported genomes of pediococci are approximately at the same level (Snauwaert *et al.,* 2015). *L. plantarum* strains usually have genomes above 3.0 Mb, probably due to their wide adaptability in a variety of foods as well as in the human gastrointestinal tract (Botta *et al.,* 2017). The strain M2011381 possessed a relatively small genome as compared to the others. Although these genomes were small, they exhibited abundant functional areas, and the proportion of such niches even accounted for more than 90% of the genome size. For the GC content of the strain M2011381, 39.8% fell into the range of the GC content of lactobacilli. The GC content may be regulated by the difference in species and the existence of plasmids. Most reported *L. plantarum* strains have higher GC content of about 44 - 45% (Illeghems *et al.,* 2015; Jeon *et al.,* 2017). A strain of *L. fermentum* was even reported to have a GC content as high as 52% (Illeghems *et al.,* 2015).

Analysis with RStudio 3.5 revealed that the five strains consisted of 840 genes that belonged to the core genome (Figure 2a). The number of singletons carried by the four strains except for *L. plantarum* WCFS1 were only estimates. *L. plantarum* WCFS1 carried the highest numbers of singletons (n = 871). The strain M2011381 represented 336 singletons. The distribution of predicted proteins into the COG functional categories of the five strains is shown in a heat map (Figure 2b). The five genomes all included more predicted proteins in categories G (carbohydrate transport and metabolism), K (transcription), E (amino acid transport and metabolism), M (cell wall/membrane/envelope biogenesis), J (translation, ribosomal structure, and biogenesis), and L (replication, recombination, and repair). Such frequencies were also observed in some *L. acidipiscis, L. salivarius,* and *L. ruminis strains* (Kazou *et al.,* 2018) because most of them relate to central cellular mechanisms. The *L. plantarum* WCFS1 chromosome contained more predicted proteins than the other four strains in the G and K categories. This strain is suggested to have a strong capacity for carbohydrate transport and metabolism, and transcription which may support the survival of *L. plantarum* strains in a range of environmental niches. The strain M2011381 possessed predicted protein frequencies approximating those of *L. brevis* ATCC 367, which was the closest neighbour of the four reference strains. However, the strain M2011381 was more adapted for amino acid transport and metabolism than *L. brevis* ATCC 367 due to the higher distribution in the E COG category. The results of the distribution of proteins into the COG functional categories of the five strains coincide with that of the phylogenetic analysis. COG functional classification of the singletons is shown in Figure 3. *L. plantarum* still displays more proteins than the other four because it had the

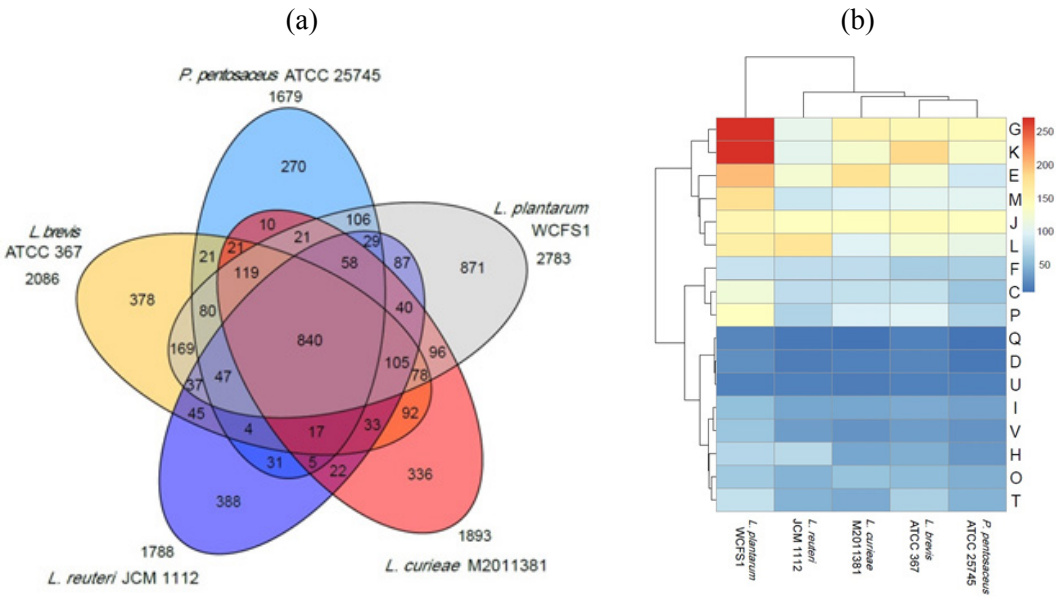

Figure 2. (a) CDS Venn diagram of the five strains. In the intersection of five strains are total core-genome. At the intersection of each pair of strains present the corresponding core-genome. The presented singleton of each strain was calculated with orthoMCL. (b) COG frequency heat map based on a two-dimensional hierarchical clustering. The horizontal axis shows the five strains, and the vertical axis shows the percentage frequency of proteins involved in each functional COG category. The letters represent the following: (G) = Carbohydrate transport and metabolism; (K) = Transcription; (E) = Amino acid transport and metabolism; (M) = Cell wall/membrane/envelope biogenesis; (J) = Translation, ribosomal structure and biogenesis; (L) = Replication, recombination, and repair; (F) = Nucleotide transport and metabolism; (C) = Energy production and conversion; (P) = Inorganic ion transport and metabolism; (Q) = Secondary metabolites biosynthesis, transport, and catabolism; (D) = Cell cycle control, cell division, and chromosome partitioning; (U) = Intracellular trafficking, secretion, and vesicular transport; (I) = Lipid transport and metabolism; (V) = Defence mechanisms; (H) = Coenzyme transport and metabolism; (O) = Posttranslational modification, protein turnover, and chaperones; and (T) = Signal transduction mechanisms.

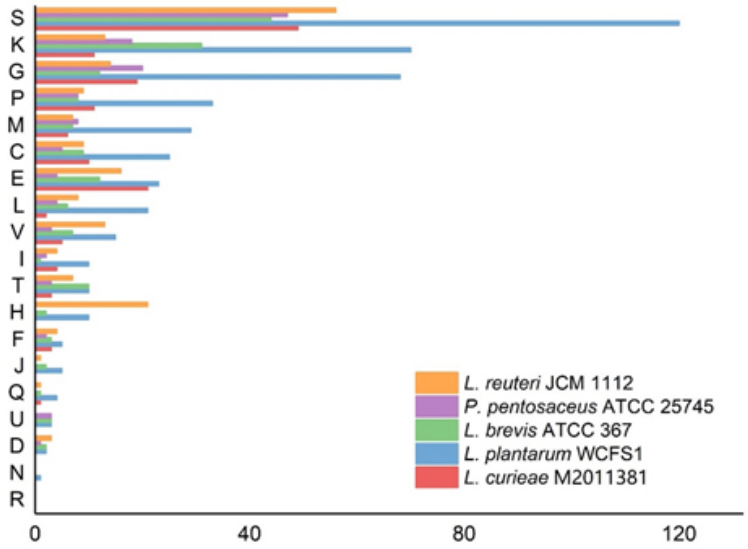

Figure 3. Distribution of singletons in COG functional categories of the five strains. The letters represent the following: (S) = Function unknown; (K) = Transcription; (G) = Carbohydrate transport and metabolism; (P) = Inorganic ion transport and metabolism; (M) = Cell wall/membrane/envelope biogenesis; (C) = Energy production and conversion; (E) = Amino acid transport and metabolism; (L) = Replication, recombination and repair; (V) = Defence mechanisms; (I) = Lipid transport and metabolism; (T) = Signal transduction mechanisms; (H) = Coenzyme transport and metabolism; (F) = Nucleotide transport and metabolism; (J) = Translation, ribosomal structure, and biogenesis; (Q) = Secondary metabolites biosynthesis, transport, and catabolism; (U) = Intracellular trafficking, secretion, and vesicular transport; (D) = Cell cycle control, cell division, and chromosome partitioning; (N) = Cell motility; and (R) = General function prediction only.

largest number of singletons among the five. The protein with majority of the five strains were K, G, P (inorganic ion transport and metabolism), and M. The superiority of proteins in the K and G COG categories appeared not only in the whole genome but also in singletons. The singletons of the strain M2011381 were distributed in most categories except H (coenzyme transport and metabolism), J, U (intracellular trafficking, secretion, and vesicular transport), D (cell cycle control, cell division, and chromosome partitioning), and N (cell motility), but the E COG category held advantageous position.

Proteolytic system

An overview of the components of the proteolytic system identified in the strain M2011381 and the four reference strains is given in Table 2. All the five strains lacked cell envelope proteinase (CEP). CEP is a serine protease with a large multi-domain structure that belongs to the subtilisin family. It is anchored to the cell wall, and initiates the casein utilisation of LAB by the extracellular degradation of casein into oligopeptides (Liu *et al.,* 2010). The deficiency of such an enzyme in the five strains coincided with their non-starter nature in bovine milk. PrtM, a membrane-bound lipoprotein, was shown to be essential for autocatalytic maturation of PrtP in *L. lactis* and *L. paracasei* (Haandrikman *et al.,* 1991; Liu *et al.,* 2010). Ancestral gene loss and metabolic simplification are central trends in the evolution of LAB. Major gene loss already occurs at the stage of the common ancestor of lactobacilli, which indicates its adaptation to nutritionally rich environments (Makarova *et al.,* 2006). The strain M2011381 might have lost this protease due to its adaptation in soy material, a nutritious niche. This genetic loss may interpret the extremely weak growth and proteolysis of the strain M2011381 in cow milk, as indicated in Figure 1c.

The second step in casein utilisation is the transportation of oligopeptides generated by CEP into the cell by the action of the oligopeptide transporter (Opp) system (Doeven *et al.,* 2005). The Opp proteins belong to a superfamily of highly conserved ATP-binding cassette (ABC) transporters that mediate the uptake of casein-derived oligopeptides. The Opp system is composed of five proteins: an oligopeptide-binding protein (OppA), two integral membrane proteins (OppB and OppC), and two nucleotide-binding proteins (OppD and OppF) (Higgins, 1992). A proton motive force (PMF)-driven dipeptide/tripeptide transporter system, DtpT, is a secondary transporter that belongs to the PTR family of peptide transporters. DtpT prefers more hydrophilic and charged di- and tripeptides (Tynkkynen *et al.,* 1993). The third system which is dependent on ATP or the energy-rich phosphorylated intermediate, Dpp (previously referred to as DtpP), transports di-, tri-, and tetrapeptides containing relatively hydrophobic branched-chain amino acids and displays the highest affinity for tripeptides. The strain M2011381 possessed both Opp and DtpT systems. Notably, the Opp system comprises all kinds of proteins which indicates that this strain specialises in transporting oligopeptides. However, the missing Dpp system implies its weaker ability in transporting hydrophobic oligopeptides, implying that the strain needs to synthesise hydrophobic amino acids for its growth. Only *L. brevis* ATCC 367 had all three known peptide transport systems, meanwhile *L. reuteri* JCM 1112 only contained DtpT. All five strains owned DtpT, indicating that this transporter is a basic need for LAB metabolism.

After the peptides are taken up into cells, they are degraded by peptidases with differing and partly overlapping specificities (Sanz *et al.,* 2001). The peptidases genetically characterised in LAB

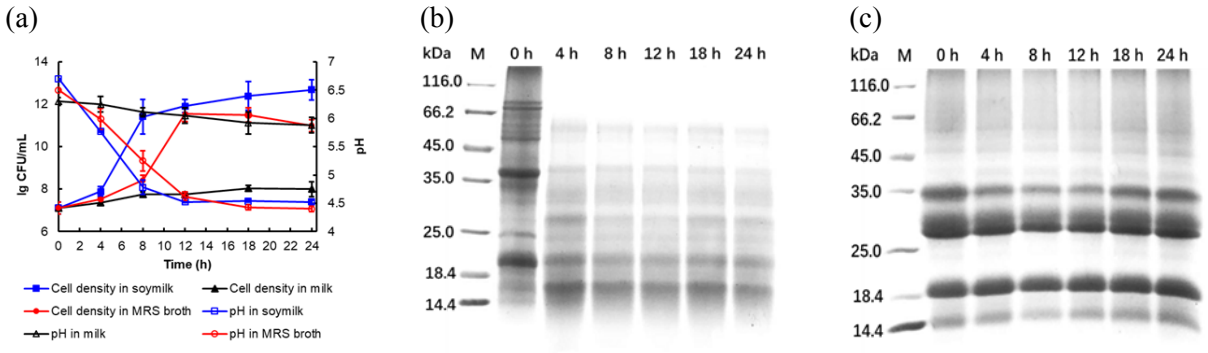

Figure 1. *L. curieae* M2011381 grows in bovine milk, soymilk, and MRS broth. (a) The cell count and pH change during the fermentation of bovine milk, soymilk, and MRS broth, respectively. (b) SDS-PAGE patterns of soymilk hydrolysates during the fermentation. (c) SDS-PAGE patterns of bovine milk hydrolysates during the fermentation. Lane 1 = molecular mass markers (M); and Lanes 2 - 7 = proteolytic activity at different times during the fermentation, respectively.

Table 2. Potential peptidases of proteolytic system in the genome of five subject LAB strains.

Peptidase	*L. curieae* CCTCC M2011381	*P. pentosaceus* ATCC 25745	*L. reuteri* JCM 1112	*L. plantarum* WCFS1	*L. brevis* ATCC 367
Aminopeptidase					
PepC	1	1	1	1	1
PepN	1	1	1	1	1
Unique aminopeptidase					
PepM	1	1	1	1	2
PepA	0	0	0	0	0
Pcp	0	0	0	0	1
Endopeptidase					
PepE/PepG	4	0	1	1	1
PepO	1	1	1	1	1
PepF	1	1	0	2	2
Dipeptidase					
PepD	2	4	5	4	5
PepV	1	1	2	1	2
Tripeptidase					
PepT	1	1	1	1	1
Proline peptidase					
PepX	1	1	1	1	1
PepI	0	0	1	1	1
PepR	0	1	1	1	1
PepL	0	0	0	1	0
PepP	0	0	1	1	1
PepQ	1	1	1	1	1

include aminopeptidase, unique aminopeptidase, endopeptidase, tripeptidase, dipeptidase, and proline-specific peptidase (Savijoki *et al.,* 2006; Liu *et al.,* 2010). Every strain of the five contained at least one enzyme of each type of the peptidases, which may be due to the hydrolysis of oligopeptides that is required for cells to get amino acids for their needs. The strain M2011381 has the same PepC, PepN, and X-prolyl dipeptidyl aminopeptidase (PepX) as the other four. Although PepC and PepN were reported as general aminopeptidases, together with PepX as the first enzymes to act on casein-derived oligopeptides (Savijoki *et al.,* 2006), their presence in all five strains indicates their importance in general protein metabolism. PepM, a methionyl aminopeptidase cleaving N-terminal methionine from proteins (Savijoki *et al.,* 2006) was also included in the strain M2011381. However, PepA or glutamyl aminopeptidase that liberates N-terminal acidic residues from 3- to 9- residue-long peptides (Liu *et al.,* 2010) was not found in the strain. PepA in *L. lactis* has been investigated for its function in milk, and it results in acidification (I'Anson *et al.,* 1995). Accordingly, defective PepA leads to the non-starter characteristic of the five strains.

The strain M2011381 had all types of endopeptidases, and PepE/PepG were the most abundant. The rich endopeptidase content and the reservation of di/tri-peptidases of the strain M2011381 can remedy its deficiency of the Dpp system. PepO with cleavage specificity on αs1-casein f1-23 and on post-proline residues of β-casein f203-209 was previously found in a non-starter strains, while PepO2 was reported from starter strains (Kunji *et al.,* 1996). The strain M2011381 and the four reference strains all had one PepO instead of PepO2, which confirms their non-starter property. PepF cleaves oligopeptides from 7- to 17-residue-long. It was found to be important for protein turnover under conditions of nitrogen starvation in *L. lactis* (Chen *et al.,* 2003). Except for

L. reuteri JCM 1112, PepF was present in all the other strains.

Many of the peptidases seem to be essential for bacterial growth or survival as they were encoded in all LAB genomes. For instance, aminopeptidases PepC, PepN, and PepM, and proline peptidases PepX and PepQ were present in all genomes, usually with the frequency of one gene per genome (Haandrikman *et al.,* 1991). However, the strain M2011381 had very few proline peptidases probably because of the absence of the CEP system, which results in no peptides comprising proline derived from casein (β-casein contains about 15.2% proline) to be transported into cell. PepP, PepI, PepL, and PepR are not essential for the growth of LAB (Liu *et al.,* 2010). Therefore, less proline content in soy protein may result in some proline peptidases to be gradually deleted during evolution in the soy environment.

Although the strain M2011381 is a non-starter in bovine milk, it has a strong capacity of fermenting plant foods such as soymilk, soy protein isolate beverage, and ginkgo nut beverage. Furthermore, the fermented materials show potent inhibitory activity of 3-hydroxy-3-methylglutaryl-coenzyme A reductase and angiotensin-I-converting enzyme because of the peptides generated during the fermentation (Liu *et al.,* 2020). These results verified that the strain possessed a specific proteolytic system due to its evolutionary environment and could play a role in fermentation of plant foods such as soymilk.

Biosynthesis pathway of amino acids

Amino acids are precursors of compounds for their characteristic flavour for food products, as well as precursors of compounds such as biogenic amines that have the potential to affect the health of consumers. The biosynthesis of amino acids represents an essential resource for LAB. The amino acids play a number of physiological roles such as intracellular pH control, generation of metabolic energy or redox power, and resistance to stress. Most amino acids originate from proteolysis and some are biosynthetically generated by lactobacilli themselves (De Angelis *et al.,* 2016). The strain M2011381 possessed the most abundant biosynthetic pathways of 12 kinds of amino acids as *L. plantarum* WCFS1. *L. plantarum* is usually recognised for its biosynthetic capacity, and contains almost all of the amino acids, thus indicating that it can survive well in many environments (Makarova et al., 2006). In the present work, L. plantarum WCFS1 was found to carry a relatively large chromosome genome, approximately 3.3 Mb, whereas the strain M2011381 had a smaller genome but with all the necessary amino acids, thus suggesting that the genes involved in amino acid synthesis are essential for the survival in the stinky tofu brine.

Conclusion

We introduced the complete sequence of the strain *L. curieae* M2011381 based on comparative genomics. The strain was revealed to have a number of genetic adaptations in the soy environment through acquired peptidases and amino acids synthetic pathways. However, the gene decay of casein utilisation was observed in M2011381 which verified that the strain is a non-starter of dairy material. The present work provides insight into the evolution of a novel LAB strain in soy environment, and indicates its use in the fermentation of soy material.

Acknowledgement

The present work was financially supported by the Opening Project of Shanghai Key Laboratory of New Drug Design, China (grant no.: 17DZ2271000), and Open Funding Project of Key Laboratory of Fermentation Engineering (Ministry of Education, China).

References

Aguirre, L., Garro, M. S., and de Giori, G. S. 2008. Enzymatic hydrolysis of soybean protein using lactic acid bacteria. Food Chemistry 111(4): 976-982.

Aguirre, L., Hebert, E. M., Garro, M. S. and de Giori, G. S. 2014. Proteolytic activity of *Lactobacillus* strains on soybean proteins. LWT - Food Science and Technology 59(2): 780-785.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. 1990. Basic local alignment search tool. Journal of Molecular Biology 215(3): 403-410.

Arndt, D., Grant, J., Marcu, A., Sajed, T., Pon, A., Liang, Y. and Wishart, D. S. 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Research 44(W1): W16-W21.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., … and Zagnitko, O. 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9: article no. 75.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., … and Yeh, L. S. L. 2005. The universal protein resource (UniProt). Nucleic Acids Research 33: D154-D159.

Botta, C., Acquadro, A., Greppi, A., Barchi, L., Bertolino, M., Cocolin, L. and Rantsiou, K. 2017. Genomic assessment in *Lactobacillus plantarum* links the butyrogenic pathway with glutamine metabolism. Scientific Reports 7: article ID 15975.

Bourdichon, F., Casaregola, S., Farrokh, C., Frisvad, J. C., Gerds, M. L., Hammes, W. P., … and Hansen, E. B. 2012. Food fermentations: microorganisms with technological beneficial use. International Journal of Food Microbiology 154(3): 87-97.

Chen, Y.-S., Christensen, J. E., Broadbent, J. R. and Steele, J. L. 2003. Identification and characterization of *Lactobacillus helveticus* PepO2, an endopeptidase with post-proline specificity. Applied and Environmental of Microbiology 69(2): 1276-1282.

De Angelis, M., Calasso, M., Cavallo, N., Di Cagno, R. and Gobbetti, M. 2016. Functional proteomics within the genus *Lactobacillus*. Proteomics 16(6): 946-962.

Diep, D. B., Godager, L., Brede, D. and Nes, I. F. 2006. Data mining and characterization of a novel pediocin-like bacteriocin system from the genome of *Pediococcus pentosaceus* ATCC 25745. Microbiology 152: 1649-1659.

Doeven, M. K., Kok, J. and Poolman, B. 2005. Specificity and selectivity determinants of peptide transport in *Lactococcus lactis* and other microorganisms. Molecular Microbiology 57(3): 640-649.

Griffiths, M. W. and Tellez, A. M. 2013. *Lactobacillus helveticus*: the proteolytic system. Frontiers in Microbiology 4: article no. 30.

Guo, T., Zhang, L., Xin, Y., Xu, Z., He, H. and Kong, J. 2017. Oxygen-inducible conversion of lactate to acetate in heterofermentative *Lactobacillus brevis* ATCC 367. Applied and Environmental Microbiology 83(21): article ID e01659-17.

Haandrikman, A. J., Kok, J. and Venema, G. 1991. Lactococcal proteinase maturation protein PrtM is a lipoprotein. Journal of Bacteriology 173(14): 4517-4525.

Higgins, C. F. 1992. ABC transporters: from microorganisms to man. Annual Review of Cell Biology 8: 67-113.

Horvath, P., Coûté-Monvoisin, A.-C., Romero, D. A., Boyaval, P., Fremaux, C. and Barrangou, R. 2009. Comparative analysis of CRISPR loci in lactic acid bacteria genomes. International Journal of Food Microbiology 131(1): 62-70.

I'Anson, K. J. A., Movahedi, S., Griffin, H. G., Gasson, M. J. and Mulholland, F. 1995. A non-essential glutamyl aminopeptidase is required for optimal growth of *Lactococcus lactis* MG1363 in milk. Microbiology 141(11): 2873-2881.

Illeghems, K., De Vuyst, L. and Weckx, S. 2015. Comparative genome analysis of the candidate functional starter culture strains *Lactobacillus fermentum* 222 and *Lactobacillus plantarum* 80 for controlled cocoa bean fermentation processes. BMC Genomics 16: article no. 766.

Jensen, L. J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T. and Bork, P. 2008. eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Research 36: D250-D254.

Jeon, S., Jung, J., Kim, K., Yoo, D., Lee, C., Kang, J., … and Cho, S. 2017. Comparative genome analysis of *Lactobacillus plantarum* GB-LP3 provides candidates of survival-related genetic factors. Infection, Genetics and Evolution 53: 218-226.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., … and Hirakawa, M. 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Research 34: D354-D357.

Kazou, M., Alexandraki, V., Blom, J., Pot, B., Tsakalidou, E. and Papadimitriou, K. 2018. Comparative genomics of *Lactobacillus acidipiscis* ACA-DC 1533 isolated from traditional Greek Kopanisti cheese against species within the *Lactobacillus salivarius* clade. Frontiers in Microbiology 9: article ID 1244.

Klaenhammer, T. R., Barrangou, R., Buck, B. L., Azcarate-Peril, M. A. and Altermann, E. 2005. Genomic features of lactic acid bacteria effecting bioprocessing and health. FEMS Microbiology Reviews 29: 393-409.

Klaenhammer, T., Altermann, E., Arigoni, F., Bolotin, A., Breidt, F., Broadbent, J., … and Siezen, R. 2002. Discovering lactic acid bacteria by genomics. Antonie von Leeuwenhoek 82(1-4): 29-58.

Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O. P., … and Siezen, R. J. 2003. Complete genome sequence of *Lactobacillus plantarum* WCFS1. Proceedings of the National Academy of Sciences of the United States of America 100(4): 1990-1995.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., … and Marra, M. A. 2009. Circos: an information aesthetic for comparative genomics. Genome Research 19(9): 1639-1645.

Kunji, E. R., Mierau, I., Hagting, A., Poolman, B. and Konings, W. N. 1996. The proteolytic systems of lactic acid bacteria. Antonie von Leeuwenhoek 70(2-4): 187-221.

Lei, X., Sun, G., Xie, J. and Wei, D. 2013. *Lactobacillus curieae* sp. nov., isolated from stinky tofu brine. International Journal of Systematic and Evolutionary Microbiology 63: 2501-2505.

Liu, M., Bayjanov, J. R., Renckens, B., Nauta, A. and Siezen, R. J. 2010. The proteolytic system of lactic acid bacteria revisited: a genomic comparison. BMC Genomics 11: article ID 36.

Liu, Y., Zhang, Y., Ro, R.-K., Li, H., Wang, L., Xie, J. and Wei, D. 2020. Gastrointestinal survival and potential bioactivities of *Lactobacillus curieae* CCTCC M2011381 in the fermentation of plant food. Process Biochemistry 88: 222-229.

Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., … and Mills, D. 2006. Comparative genomics of the lactic acid bacteria. Proceedings of the National Academy of Sciences of the United States of America 103(42): 15611-15616.

Melini, F., Melini, V., Luziatelli, F., Ficca, A. G. and Ruzzi, M. 2019. Health-promoting components in fermented foods: an up-to-date systematic review. Nutrients 11(5): article ID 1189.

Morita, H., Toh, H., Fukuda, S., Horikawa, H., Oshima, K., Suzuki, T., … and Hattori, M. 2008. Comparative genome analysis of *Lactobacillus reuteri* and *Lactobacillus fermentum* reveal a genomic island for reuterin and cobalamin production. DNA Research 15(3): 151-161.

Ramachandran, P., Lacher, D. W., Pfeiler, E. A. and Elkins, C. A. 2013. Development of a tiered multilocus sequence typing scheme for members of the *Lactobacillus acidophilus* complex. Applied and Environmental Microbiology 79(23): 7220-7228.

Ren, C., Faas, M. M. and de Vos, P. 2020. Disease managing capacities and mechanisms of host effects of lactic acid bacteria. Critical Reviews in Food Science and Nutrition (in press).

Sanz, Y., Lanfermeijer, F. C., Renault, P., Bolotin, A., Konings, W. N. and Poolman, B. 2001. Genetic and functional characterization of *dpp* genes encoding a dipeptide transport system in *Lactococcus lactis*. Archives of Microbiology 175(5): 334-343.

Savijoki, K., Ingmer, H. and Varmanen, P. 2006. Proteolytic systems of lactic acid bacteria. Applied Microbiology and Biotechnology 71(4): 394-406.

Snauwaert, I., Stragier, P., de Vuyst, L. and Vandamme, P. 2015. Comparative genome analysis of *Pediococcus damnosus* LMG 28219, a strain well-adapted to the beer environment. BMC Genomics 16(1): article no. 267.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28(10): 2731-2739.

Tsuda, H. and Shibata, E. 2017. Bioconversion of daidzin to daidzein by lactic acid bacteria in fermented soymilk. Food Science and Technology Research 23(1): 157-162.

Tynkkynen, S., Buist, G., Kunji, E., Kok, J., Poolman, B., Venema, G. and Haandrikman, A. 1993. Genetic and biochemical characterization of the oligopeptide transport system of *Lactococcus lactis*. Journal of Bacteriology 175(23): 7523-7532.

Wang, Y., Wang, Y., Lang, C., Wei, D., Xu, P. and Xie, J. 2015. Genome sequence of *Lactobacillus curieae* CCTCC M2011381^{T}, a novel producer of gamma-aminobutyric acid. Genome Announcements 3(3): article ID e00552-15.

Xie, C.-L., Hwang, C. E., Oh, C. K., Yoon, N. A., Ryu, J. H., Jeong, J. Y., … and Lee, D. H. 2017. Fermented soy-powder milk with *Lactobacillus Plantarum* P1201 protects against high-fat diet-induced obesity. International Journal of Food Science and Technology 52(7): 1614-1622.

Yang, S.-J., Lee, J.-E., Lim, S.-M., Kim, Y.-J., Lee, N.-K. and Paik, H.-D. 2019. Antioxidant and immune-enhancing effects of probiotic *Lactobacillus plantarum* 200655 isolated from kimchi. Food Science and Biotechnology 28: 491-499.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. and Wishart, D. S. 2011. PHAST: a fast phage search tool. Nucleic Acids Research 39: W347-W352.