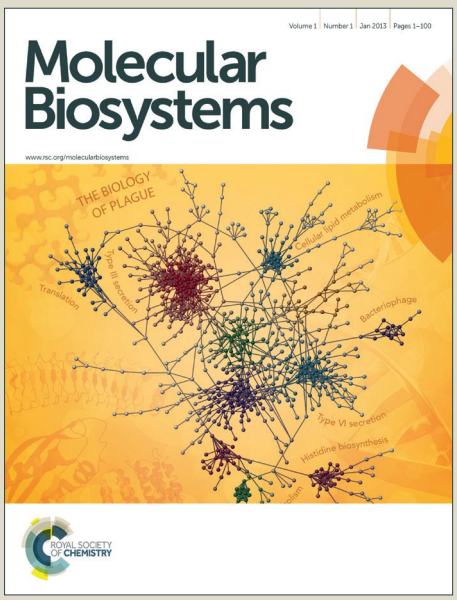
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Molecular BioSystems

ARTICLE

Strand-specific RNA-seq analysis of the *Lactobacillus delbrueckii subsp. bulgaricus* transcriptome

Received 00th July 2015,
Accepted 00th January 2015

DOI: 10.1039/x0xx00000x

www.rsc.org/

Huajun Zheng[a]*, Enuo Liu[a]*, Tao Shi[a], Luyi Ye[a], Tomonobu Konno[b], Munehiro Oda[c] and Zai-Si Ji[a, b]&

Lactobacillus delbrueckii subsp. bulgaricus 2038 (*Lb. bulgaricus* 2038) is an industrial bacterium that is used as a starter for dairy products. We proposed several hypotheses concerning its industrial features previously. Here, we utilized RNA-seq to explore the transcriptome of *Lb. bulgaricus* 2038 from four different growth phases in whey condition. The most abundantly expressed genes in the four stages were mainly involved in translation (for logarithmic stage), glycolysis (for control/lag stages), lactic acid production (all the four stages), and 10-formyltetrahydrofolate production (for stationary stage). The high expression of genes like D-lactate dehydrogenase was thought as a result of energy production, and consistent expression of EPS synthesis genes, restriction-modification (RM) system and CRISPR/Cas system were validated for explaining the advantage of this strain in yoghourt production. Several postulations, like NADPH production through GapN bypass, converting aspartate into carbon-skeleton intermediates, and formate production through degrading GTP, was proved not working in this culture condition. The high expression of helicase genes and co-expressed amino acids/oligopeptides transporting proteins indicated the helicase might mediate the strain obtaining nitrogen source from the environment. Transport system of *Lb.bulgaricus* 2038 was found regulated by antisense RNA, hinting the potential application of non-coding RNA in regulating lactic acid bacteria (LAB) gene expression. Our study has primarily uncovered *Lb. bulgaricus* 2038 transcriptome, which could gain a better understanding of the regulation system in Lb. bulgaricus and promote its industrial application.

Introduction

Lactobacillus delbrueckii subsp. *bulgaricus* (*Lb. bulgaricus*) is traditionally used in the production of fermented foods, beverages, and feed. In addition to lactic acid production, *Lb. bulgaricus* is useful for protein hydrolysis and the formation of numerous compounds, such as flavors and peptides [1, 2]. The industrial strain *Lb. bulgaricus* 2038 is well suited for the large-scale production of yoghurt, and the genetic basis associated with its industrial features was elucidated through genome analysis (i.e., transforming aspartate into carbon-skeleton intermediates, lysine biosynthesis, formate production, and acid tolerance) [3]. Through the use of microarray analysis, we have constructed an amino acid acquisition model of *Lb. bulgaricus* 2038, which revealed that this strain uses different intracellular peptidases when grown in casein or whey as the sole nitrogen source [4-6]. Sieuwerts *et al*. described the transcriptome profiles of a mixed-culture of *Streptococcus thermophilus* and *Lb. bulgaricus* in cow milk using microarray analysis and demonstrated that the proteolytic system of *Lb. bulgaricus* could not liberate sufficient branched-chain amino acids (BCAA) and sulphur amino acids [7]. However, other features in addition to nitrogen usage in *Lb. bulgaricus* have not been validated through gene expression profiling. *Lb. bulgaricus* 2038 entered logarithmic stage growth after two hours in casein culture condition and the whole life cycle of this microorganism takes nine hours [6], while in whey the bacteria enters logarithmic stage after 20 minutes and completes its life cycle in three hours [5]. On the other hand, the rapid growth of *Lb. bulgaricus* 2038 in whey (ph5.2) is more like its growth in industrial condition, so the idea to study the genetic basis associated with industrial features of

[a] *Laboratory of Medical Foods, Shanghai Institute of Planned Parenthood Research, 2140 Xie-Tu Road, Shanghai 200032, China.*
[b] *Division of Research and Development, Meiji Co., Ltd., 540 Naruda, Odawara, Kanagawa 250-0862, Japan.*
[c] *GraduateSchool of Bioresource Sciences, Nihon University, 1866 Kameino,Fujisawa city, Kanagawa 252-0880, Japan.*
**contributed equally to this work*
& *Correspondence: zai-si_ji@kta.biglobe.ne.jp*
†Electronic Supplementary Information (ESI) available. See
DOI: 10.1039/x0xx00000x

Lb. bulgaricus 2038 can be achieved through analyzing transcriptome profile in whey.

RNA-seq is an excellent approach for transcriptome profiling. It uses deep-sequencing technologies to directly determine the cDNA sequence and has a relatively low background compared with the use of microarrays [8]. The most striking advantage of RNA-seq is that it can provide the absolute expression value of each gene, thereby making it possible to determine the expression abundance of each gene in the same stage. Moreover, the ability to obtain highly reproducible results with few systematic differences makes it possible to sequence each sample without the requirement of technical replicates [9]. In addition to detecting gene expression and transcriptional regulation under different physiological conditions, the development of RNA-Seq technology (especially strand-specific RNA-seq) has enabled the discovery of operon structures, transcription start sites, 5′ and 3′ UTRs, and non-coding RNAs [10-13].

RNA-seq technology had been applied to the study of lactic acid bacteria. *Lactobacillus plantarum WCFS1* was selected as a model organism to validate the reliability of RNA-Seq technology compared with the use of DNA microarrays [14]; moreover, genes associated with the utilization of tetrasaccharides by *Lactobacillus ruminis* L5 were identified by RNA-seq [15]. However, no RNA-seq work had been reported for *Lb. bulgaricus*. This lack might be partly due to the high proportion of rRNA in its transcriptome. The *Lb. bulgaricus* genome contains nine rRNA operons that encode over 90% of the total RNA content, as described in our previous study (data not shown). rRNA removal and strand-specific library construction were two main problems associated with the use of RNA-seq with this bacteria, even several protocols and kits had been developed to solve these problems [11, 16, 17]. In this study, the Ribo-Zero rRNA Removal Kit was selected after several trials because this kit diminished 99% of the rRNAs from the total RNA. We took advantage of strand-specific RNA-seq technology to study the transcriptome of *Lb. bulgaricus* 2038 in the four growth stages in whey, and investigated the functions of the genes that were differentially expressed during these phases.

Materials and methods

Bacterial strains, media, and growth conditions

Lb. bulgaricus 2038 were propagated by passage through three sequential pre-cultures in a modified partially chemically defined medium (containing 0.6 % acid-hydrolyzed casein, 0.01 % asparagine, 0.01 % tryptophan, and 0.02 % cysteine as the only nitrogen source) incubated at 40°C overnight under anaerobic conditions. Then, the cultures were washed with saline solution, stock cultures were prepared in 10 % (v/v) glycerol and stored at -80°C. Glycerol stock cultures were inoculated in 8 ml of the same medium (final density 6 x 10^6 cells/ml), incubated for 5 min at 40°C with shaking (10 rpm; TAITEC Rotator RT-50; Saitama, Japan), and then collected as the start control. To isolate total RNA during different growth phases, tubes containing 8 ml of whey medium were inoculated with the glycerol stock culture (final density 6 x 10^6 cells/ml) at 40°C with shaking (10 rpm) under the same conditions used for the start control. Cells were collected after 20, 120, and 270 min. At each collection point, 1 ml of ice-cold ethanol/phenol stop solution [5% water-saturated phenol (pH<7.0) in ethanol] was added to each tube; then, the tubes were centrifuged at 9,000xg at room temperature for 1 min. The cells were frozen with liquid nitrogen and stored at -80°C. The whey medium was prepared with 0.024 g MgSO4, 0.003 g MnSO4, 0.001 g FeSO4, 0.100 g Tween 80, 20μg P-amino benzoic acid, 10μg folic acid, and 89 ml of whey solution; the pH was adjusted to 5.20 with NaOH at 20°C, and the solution was deoxygenated overnight with an AnaeroPack Anaero (Mitsubishi Gas Chemical America,Inc., New York, NY, USA)prior to use. All reagents were filter-sterilized separately. The whey solution was made from heat-treated (90°C for 15 s), defatted, and concentrated fresh Holstein cow milk; the pH was adjusted to 4.6 with 20 % HCl at 20°C and centrifuged for 10 min at 1,000xg at 20°C. Approximately 700 ml of the supernatant (whey solution) was obtained and filter-sterilized.

RNA isolation and removal of rRNA

Frozen cells were resuspended in 0.5 ml of buffer containing 0.1 M LiCl, 0.01 M EDTA, 0.01 M Tris–HCl, and 1 % SDS, then vortexed two times for 45 s at speed 6.5 in a Thermo Savant FastPrep FP100A/BIO101 homogenizer. Total RNA was isolated using the acid–phenol method [18] and treated with RNase-free DNase (TAKARA BIO INC., Shiga, Japan) at 37°C for 1 hr. RNA quality and quantity were assessed using a Thermo Scientific Nanodrop 2000 and Agilent 2100 bioanalyzer, respectively. Then, the RNA samples were purified using the RNeasy MinElute Cleanup Kit (Qiagen, Hilden, Germany) prior to the rRNA

removal step. To remove rRNA, the Ribo-Zero rRNA Removal Kit (Epicentre, Madison, WI, USA) was applied according to manufacturer's instructions.

RNA-seq

Construction of strand-specific cDNA libraries was performed using the RNA transcriptome discovery Kit (Gnomegen, San Diego, CA, USA), which ligated an Illumina compatible 5' adaptor to the 5' end of heat fragmented mRNA, following the manufacturer's instructions. cDNA with a range between 100bp and 550bp was obtained by gel extraction followed by amplification using the TruSeq PE Cluster Kit (Illumina, San Diego, CA, USA). Amplified cDNA fragments were pair-end sequenced using Illumina sequencing technology (Illumina High-seq 2000) with the 1st read from the 5' end of the transcripts. The sequencing data were submitted to the National Center for Biotechnology Information Sequence Read Archive under Accession No. SRP049278.

Analysis of gene expression levels

After adaptor trimming and quality trimming, the clean reads were mapped to the *Lb. bulgaricus 2038* genome using Bowtie2 [19]. The read numbers of each gene were first transformed into RPKM (Reads Per Kilo bases per Million reads) [20], then differentially expressed genes were identified with the DEGseq package using the MARS method (MA-plot-based method with Random Sampling model) [21]. We defined genes with at least a 2-fold change between two samples and an FDR (false discovery rate) less than 0.001 as differentially expressed genes. Operon structures and non-coding RNA were predicted using Rockhopper [22]. An average linkage hierarchical clustering was performed with Genesis 1.7.6 [23] using genes with RPKM values greater than 10 in at least one of the four stages; the RPKM values were log 2 transformed. The sRNA target prediction was performed using RNAPredator [24]. CRISPRs in the genome were detected using CRISPRFinder [25]. Spearman correlation coefficient between two variables was calculated using R command 'cor.test'. For example, when comparing the gene expression level of 18 genes in two stages, the variable 1 was gene expression value of the 18 genes in stage1, and variable 2 was gene expression value of the 18 genes in stage2.

Quantitative real-time PCR (qRT-PCR) verification

Total RNA was treated with the DNase I and quantitative RT-PCR was performed on RNA samples to confirm the absence of DNA contamination using 16S primers. The cDNA synthesis was performed using AMV First Strand cDNA Synthesis Kit (Sangon Biotech, Shanghai, China) following the manufacturer's instruction. The genes and primers used for qRT-PCR are shown in **Table S1**. qRT-PCR was performed with the ABI StepOne system using the Fast SYBR Green Master Mix (Thermo Fisher Scientific, Rockland, IL, USA). The sigma A homologue gene *LBU_1062* was used as an internal control [26]. The relative gene expression was calculated from three independent experiments and analyzed using the $2^{-\Delta\Delta CT}$ (where *CT* is cycle threshold) value method [27].

Results and discussion

The transcriptome of *Lb. bulgaricus* 2038

To investigate the transcriptome of *Lb. bulgaricus* 2038 by RNA-seq during fermentation, we harvested this bacteria cultures growing in whey medium and selected the checkpoints at 5 min (control), 20 min (lag phase), 120 min (logarithmic phase), and 270 min (stationary phase) according to the bacteria's growth curve (**Figure 1)**. A total of 7.9 Gb of 2x100 bp data was obtained

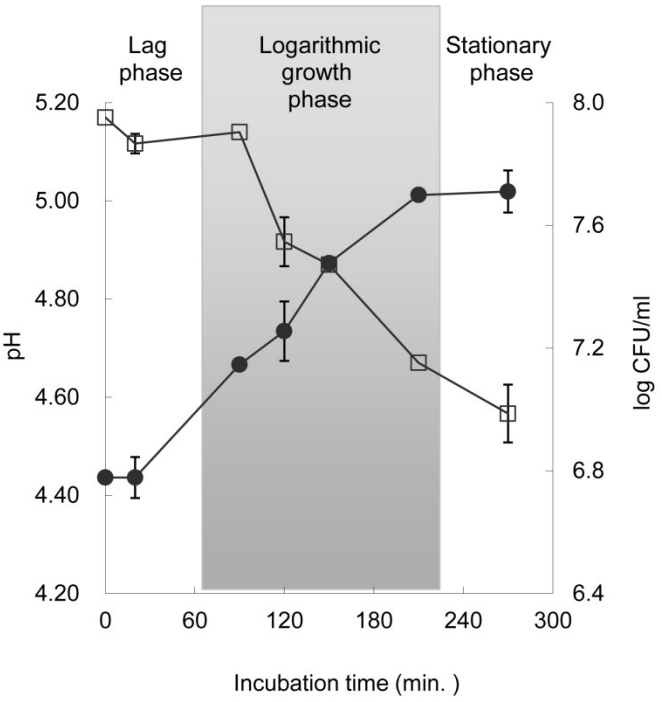


Figure 1. Growth and pH changes in *L. bulgaricus* 2038 culture at 40°C in whey medium (● log CFU/ml, □ pH). Results are expressed as the mean of three independent experiments; error bars represent standard deviation. Cultures were separately collected at checkpoints 5 min (control), 20 min (lag phase), 120 min (logarithmic phase), and 270 min (stationary phase).

Molecular BioSystems

ARTICLE

Table 1. Sequence statistics of *Lb.bulgaricus* 2038

	control (5 min)	lag (20 min)	logarithmic (120 min)	stationary (270 min)
raw reads (pair)	3,548,412	10,508,406	15,015,465	11,747,332
clean reads (pair)	3,487,515	10,172,967	14,560,162	11,458,290
Reads mapped to Genome	3,456,790 (99.12%)	10,109,700 (99.38%)	14,405,030 (98.93%)	11,416,239 (99.63%)
Reads mapped to Ribosomal RNA	7,208 (0.21%)	33,050 (0.32%)	56,185 (0.39%)	110,353 (0.96%)
Reads mapped to tRNA	17,032 (0.49%)	9,039 (0.09%)	13,333 (0.09%)	2,318 (0.02%)
Reads mapped to Genes	212,3271 (60.88%)	8,621,799 (84.75%)	13,429,280 (92.23%)	10,539,181 (91.98%)
antisense mapping to gene	24,524 (0.69%)	134878 (1.33%)	133,645 (0.92%)	93,936 (0.8%)
Reads mapped to Intergenic Region	1,284,755 (36.84%)	1,310,934 (12.89%)	772,587 (5.31%)	670,451 (5.87%)
Mapped Genes	1,757	1,772	1,788	1,779

after removing adaptors and low-quality bases, with 3.5 million pairs of reads for the control phase, 10.2 million for the lag phase, 14.6 million for the logarithmic phase, and 11.5 million for the stationary phase (**Table 1**). The results represented all of the 1,790 genes of *Lb. bulgaricus* 2038 [3], and saturation curves and gene coverage indicated a completely adequate sequencing depth (**Figure S1, S2**).

The mean RPKM (Reads Per Kilo bases per Million reads) was 842 in the control, 711 in the lag phase, 731 in the logarithmic phase, and 619 in the stationary phase (**Table S2**). Gene's high and low expression was usually defined by arbitrary threshold [28, 29], here we define genes with RPKM<10 as low expression genes. A total of 55 genes were found to have low expression levels in each stage (RPKM<10); 33 of these genes were pseudogenes, and eight were hypothetical proteins (**Table S2**). The other 14 functional genes with low expression levels might indicate that they were silenced when grown in whey.

To study the gene expression changes of *Lb. bulgaricus 2038* during fermentation, we set our criteria as a fold change>2 and FDR<0.001 (**Table S2**). This strategy revealed 402 significantly regulated genes in the lag phase relative to the control, 821 significantly regulated genes in the logarithmic phase relative to the lag phase, and 587 significantly regulated genes in the stationary phase relative to the logarithmic phase (**Table S3**). Strikingly, we found five genes that were consecutively down-regulated throughout the entire growth process (**Table S4**), indicating their potential role in the initiation of growth. In contrast, no consecutively up-regulated genes were observed during the growth process.

The codon adaptation index (CAI) had been widely used to assess protein expression levels [30, 31] by comparing the codon usage bias of a gene with a reference set of highly expressed

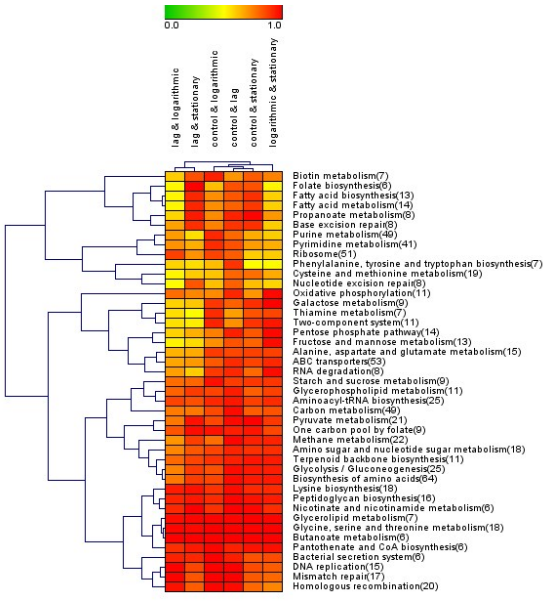


Figure 2. Heatmap representing correlation coefficients of genes involved in same pathway between different growing phases. The spearman correlation coefficients (R value) was represented by color, with red indicating high correlation (R=1) and green indicating the weak correlation (R=0). The number in bracket was the gene number involved in pathway.

Molecular BioSystems

ARTICLE

Table. 2 The most abundantly expressed gens in each stage of *Lb.bulgaricus* 2038

		control			lag			logarithmic			stationary			CAI	
Gene	**Product**	**RPKM**	**proportion**	**rank***	**RPKM**	**proportion**	**rank**	**RPKM**	**proportion**	**rank**	**RPKM**	**proportion**	**rank**	**Value**	**Order**
LBU_1429	conserved hypothetical protein	251073	16.7%	1	46097	3.6%	3	6273	0.5%		8424	0.8%		0.306	568
LBU_0225	small heat shock protein	86529	5.7%	2	71607	5.6%	1	3613	0.3%		4858	0.4%		0.365	248
LBU_1379	chaperonin GroES	38454	2.6%	3	71071	5.6%	2	7463	0.6%		8252	0.7%		0.382	193
LBU_1747	conserved hypothetical protein	33653	2.2%	4	15254	1.2%	6	4908	0.4%		15516	1.4%		0.234	1324
LBU_1378	chaperonin GroEL	19626	1.3%	5	36716	2.9%	4	4178	0.3%		4201	0.4%		0.427	106
LBU_1106	enolase	17814	1.2%	6	14706	1.2%	7	9122	0.7%		18304	1.7%	6	0.543	5
LBU_0066	D-lactate dehydrogenase	17643	1.2%	7	19349	1.5%	5	8753	0.7%		23143	2.1%	2	0.541	6
LBU_0532	glyceraldehyde 3-phosphate dehydrogenase	14632	1.0%	8	9010	0.7%		16512	1.3%	7	17033	1.5%	9	0.507	23
LBU_0669	elongation factor Tu	14170	0.9%	9	11974	0.9%	9	23816	1.8%	2	21873	2.0%	4	0.634	1
LBU_0031	4-oxalocrotonate tautomerase	11142	0.7%	10	2548	0.2%		910	0.1%		867	0.1%		0.354	291
LBU_0064	cobalamin adenosyltransferase	6656	0.4%		6004	0.5%		6647	0.5%		18006	1.6%	8	0.234	1314
LBU_0320	ribosomal protein L3	6284	0.4%		7826	0.6%		15890	1.2%	8	2699	0.2%		0.433	95
LBU_0328	ribosomal protein L29	5630	0.4%		4831	0.4%		19641	1.5%	4	3493	0.3%		0.55	4
LBU_0329	ribosomal protein S17	6785	0.5%		4043	0.3%		15141	1.2%	10	2192	0.2%		0.538	10
LBU_0330	ribosomal protein L14	9286	0.6%		6324	0.5%		21224	1.6%	3	3219	0.3%		0.384	189
LBU_0331	50S ribosomal protein L24	9531	0.6%		4404	0.3%		18367	1.4%	6	3331	0.3%		0.409	132
LBU_0341	adenylate kinase	5465	0.4%		2978	0.2%		15671	1.2%	9	2221	0.2%		0.381	199
LBU_0342	translation initiation factor IF-1	10285	0.7%		12691	1.0%	8	31986	2.4%	1	8311	0.8%		0.382	191
LBU_0792	conserved hypothetical protein	6350	0.4%		11273	0.9%	10	1505	0.1%		2402	0.2%		0.25	1116
LBU_0869	putative formate-tetrahydrofolate ligase	3725	0.2%		2175	0.2%		5028	0.4%		25338	2.3%	1	0.302	598
LBU_0870	putative prolipoprotein signal peptidase	2905	0.2%		1790	0.1%		4207	0.3%		18196	1.6%	7	0.256	1041
LBU_1420	ribosomal protein L11	9433	0.6%		10599	0.8%		18765	1.4%	5	22842	2.1%	3	0.473	39
LBU_1511	putative oxalate:formate antiporter	1369	0.1%		387	0.0%		2356	0.2%		20648	1.9%	5	0.281	774
LBU_1638	conserved hypothetical protein	3322	0.2%		3259	0.3%		8095	0.6%		16976	1.5%	10	0.396	164

rank* represent expression order of the gene in whole genome, e.g., 1 means the most abundantly expressed gene

Molecular BioSystems

ARTICLE

genes [32]. The CAI values of *Lb. bulgaricus* 2038 genes were previously calculated [3], and the correlation coefficients were found to be relatively low between the CAI values and the real gene expression values (RPKM) using the Spearman Test (R=0.17 in the control, R=0.25 in the lag stage, R=0.54 in the logarithmic stage, and R=0.37 in the stationary stage). This result is quite reasonable because gene expression is affected by many factors. However, five of the top ten highly expressed genes calculated by CAI were highly expressed during the four stages (**Table 2**), including the genes encoding the elongation factor Tu, enolase, and D-lactate dehydrogenase (D-LDH). This result in part supports the reasonability of the CAI calculation.

The accuracy of the transcriptome data reflecting the gene expression levels was validated by the comparison of correlation coefficients of genes involved in the same pathway between different growth phases (**Figure 2**). The gene expression in most pathways showed a high correlation. Taking the 18 genes involved in glycine, serine and threonine metabolism as an example, the expression values of the 18 genes in control phase showed very high correlations (R^2 >0.95) with those in lag, logarithmic or stationary phase. In addition, the expression profile of the top 10 highly expressed genes in each growth stage (**Table 2**) showed a high correlation coefficients (R=0.86 for the gene expression changes of stationary stage vs control) with microarray data of our previous study [5].

In addition, seven genes, including highly expressed genes in each stage, helicase gene *LBU_1514* and pyruvate oxidase gene *LBU_1788* were selected for qRT-PCR validation (**Table S1**). The qRT-PCR results displayed similar trends and a high level of correlation (R^2=0.73) with those observed in RNA-seq results (**Figure S3**), implying that the RNAseq data is reliable.

The most abundantly expressed genes during fermentation

The top ten expressed genes in each stage were investigated (**Table 2**). The most abundantly expressed gene in the control phase was *LBU_1429*, which encoded a hypothetical protein and accounted for 16.7% of the total transcripts. Based on the average linkage hierarchical clustering, *LBU_1429* clustered together with genes encoding the chaperonin GroEL, chaperonin GroES and a small heat shock protein; therefore, we propose that *LBU_1429* encodes a protein involved in protein folding, sorting and degradation. High expression of *LBU_1429* and other chaperonin genes in the initiation of growth guarantee the normal growth of bacteria through preventing irreversible protein denaturation in the new culturing environment. In the lag phase, *LBU_0225*, which encodes a small heat shock protein (sHSP), accounted for 5.6% of the whole transcriptome. The sHSP was characterized by a chaperone activity to prevent irreversible protein denaturation [33]. Its high expression in both control and lag phase ensured the function of extracellular proteases and transport systems which satisfied the growth need of the strain. Meanwhile, eight of the top expressed genes were similar between the control and lag phases, including the genes encoding two enzymes involved in glycolysis (glyceraldehyde 3-phosphate dehydrogenase and enolase), D-lactate dehydrogenase, and the chaperonins GroES/GroEL. This finding indicated that glycolysis played an important role during the initiation of the culture.

The top ten expressed genes in the logarithmic phase were definitely different from the previous two stages. Only three genes (*LBU_0532*, *LBU_0669*, and *LBU_0342*) appeared in the top ten genes of the control/lag stages (**Table 2**). *LBU_0669* encoded the translation elongation factor Tu, which aids in the interaction between the aminoacyl-tRNA and the ribosome. Because it is a key component in translation, *LBU_0669* was highly expressed during all four growth stages. *LBU_0342,* which encoded the translation initiation factor IF-1, was the most abundantly expressed gene in the logarithmic stage and accounted for 2.5% of the whole transcriptome. *LBU_0532* encoded a glyceraldehyde 3-phosphate dehydrogenase involved in glycolysis. The other seven genes encoded six ribosomal proteins and an adenylate kinase that catalyzed ADP/AMP conversion. The high expression levels of these genes are in

agreement with the fast growth of the bacteria during the logarithmic stage.

Five of the top 10 expressed genes during the stationary phase were not detected during the previous three stages. Among them, *LBU_0869* was the most abundantly expressed transcript and occupied 2.3% of the transcriptome. This gene encoded the formate-tetrahydrofolate ligase, which produces 10-formyltetrahydrofolate (10-CHO-THF) from formate. This result signifies the importance of 10-CHO-THF during the stationary stage. 10-CHO-THF acts as a donor of formyl groups during anabolism and plays an important role in purine biosynthesis. Indeed, 10-CHO-THF is a substrate for both phosphoribosyl glycinamide formyltransferase (LBU_1234, K11175, EC:2.1.2.2) and phosphoribosylaminoimidazolecarboxamide formyltransferase (LBU_1233, K00602, EC:2.1.2.3) and is involved in the formylation of the methionyl initiator tRNA (fMet-tRNA), where 10-CHO-THF is a substrate for methionyl-tRNA formyltransferase (LBU_1208, K00604, EC:2.1.2.9). 10-CHO-THF can further be transformed into 5-methyltetrahydrofolate (**Figure 3**). Additionally, *LBU_0066*, which encodes D-lactate dehydrogenase, was the 2nd most abundant transcript, coinciding with our finding that D-lactic acid was the main production in stationary phase (**Table S5**) .

Gene expression associated with industrial features

Carbon metabolism and energy generation *Lb. bulgaricus* 2038 can utilize lactose, glucose, fructose and mannose [3]. We re-constructed the transportation and metabolism pathway of carbohydrates based on the gene expression profiles (**Figure 4**). Lactose was the main carbon source for *Lb. bulgaricus* during dairy fermentation, and the lactose permease (LacS) and beta-galactosidase (LacZ) genes showed consistently high expression levels in all four growth stages. A similarly high expression level was observed in the mannose-specific PTS system (*LBU_1520-LBU1522*), which imports glucose and mannose. In contrast, the gene (*LBU_1166*) encoding the putative glucose uptake protein exhibited a low expression level, suggesting that glucose was mainly imported by the mannose-specific PTS system. The fructose-specific PTS system showed stage specificity, with high expression in the control and lag stages.

More than 90% of the pyruvate in *Lb. bulgaricus* is converted into D-lactate [34]. This finding was in agreement with the expression profile of the pyruvate metabolism genes of *Lb. bulgaricus* 2038 (**Figure 4**). These genes consisted of three D-LDH genes (*LBU_0066*, *LBU_0860* and *LBU_1637*), two L-LDH genes (*LBU_0059* and *LBU084*), and one pyruvate oxidase gene (*LBU_1788*). The expression level (RPKM) of the three D-LDH genes occupied 93% of all six genes, with *LBU_0066* showing a very high expression level in all four stages (**Table 2**).

Due to incomplete citrate cycle, substrate level phosphorylation of glycolysis was the main ATP production step of *Lb.bulgaricus*. Lactate production can regenerate NAD^+ and compensate the NAD^+ consumption in glycolysis, and thus guarantee the progress of glycolysis. So we can conclude that lactate production was a by-product of energy production. This can be supported by the fact that the D-LDH gene *LBU_0066* showed high expression level (RPKM>1,000) in all the four growth stages, similar with expression level of genes involved in glycolysis.

In addition, acetyl-phosphate transforming into acetate can also produce ATP by acetate kinase (LBU_0581) (**Figure 4**). In *Lb.bulgaricus* 2038, acetyl-phosphate could be transformed from pyruvate by pyruvate oxidase (LBU_1788) or from xylulose 5-phosphate by xylulose-5-phosphate-fructose phosphoketolase (LBU_0444). But both pathways had restrictions. Reaction of

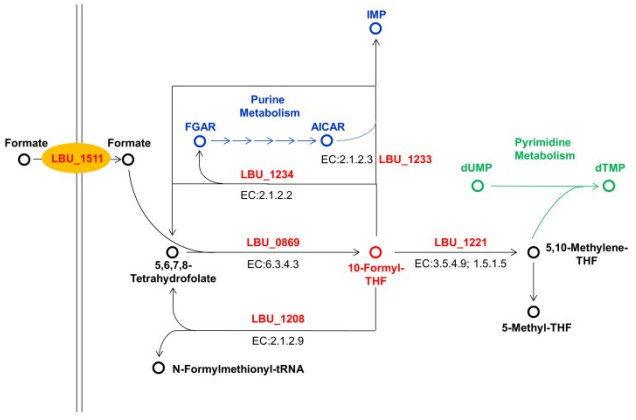


Figure 3. The role of 10-formyltetrahydrofolate (10-CHO-THF) in metabolism.

Molecular BioSystems

ARTICLE

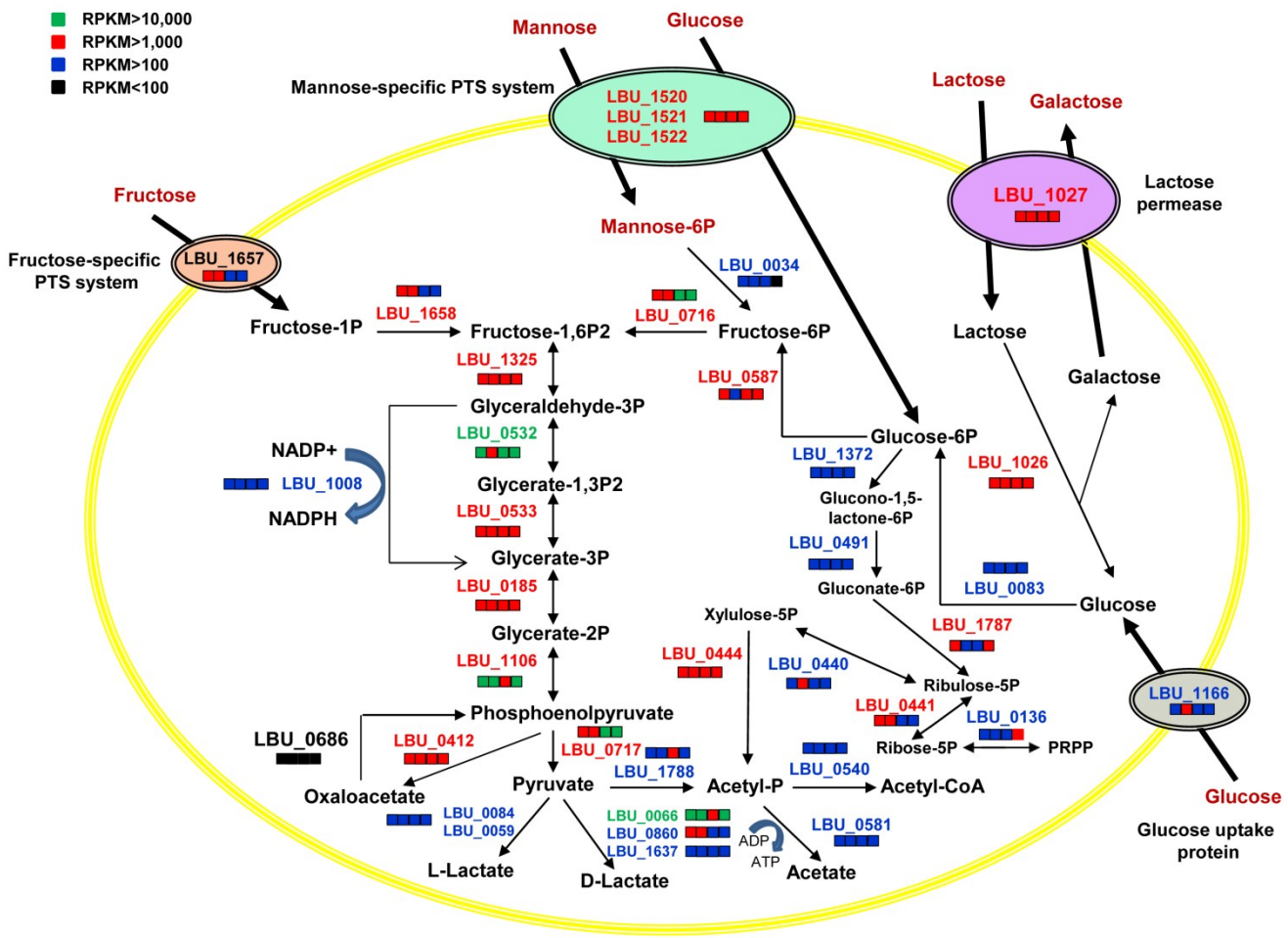


Figure 4. Carbohydrate transporting and metabolism pathway re-constructed based on gene expression profiles. Colors represented different gene expression level. Green-RPKM>10000, Red-1000<RPKM<1000, Blue-100<RPKM<1000, Black-PRKM<100.

pyruvate oxidase needs oxygen, which was rare in the *Lb.bulgaricus* culture, so xylulose 5-phosphate was the main substrate of acetyl-phosphate formation. This was supported by the fact that *LBU_0444* expressed in high level in all the four growth stages, while *LBU_1788* showed high expression only in logarithmic stage (**Table S2, Figure 4**). But xylulose 5-phosphate was the production of ribulose 5-phosphate, which could also be transformed into PRPP, the precursor of both purine and pyrimidine synthesis. So when the bacteria was in a need of nucleotide synthesis, xylulose 5-phosphate production would be restricted. Meanwhile, acetyl-phosphate was the only precursor of acetyl-CoA in *Lb.bulgaricus* 2038, so the energy production pathway from acetyl-phosphate was competed by acetyl-CoA requirement.

Because the transketolase gene was lacked in the pentose phosphate pathway (HMP), NADPH production had been postulated to be compensated by a $NADP^+$-dependent glyceraldehyde-3P dehydrogenase (GapN) [35], thereby bypassing the pathway from glyceraldehyde-3P to glycerate-3P (**Figure 4**). In *Lb. bulgaricus* 2038, the expression of the GapN gene

LBU_1008 (average PRKM=164) was greatly reduced compared to *LBU_1372* (average RPKM=412) and *LBU_1787* (average RPKM=1,122) (**Table S2**), which are the two genes that produce NADPH in HMP. The gene *LBU_0532* encoding the glyceraldehyde-3P dehydrogenase exhibited almost 100-fold higher expression than *LBU_1008*, indicating that most of the glyceraldehyde-3P was transformed into glycerate-3P through the traditional ATP production pathway instead of the GapN bypass. Therefore, we speculated that the contribution of GapN to NADPH production was less than anticipated, or the role of GapN was exerted only in specific circumstances.

Degenerate Amino Acid synthesis In silico analysis revealed that *Lb. bulgaricus* 2038 could synthesize lysine, threonine, and asparagine from aspartate [3, 4]. However, the enzymes that transform oxaloacetate into aspartate were not revealed through KAAS analysis [36]. Our previous study postulated that aspartate could be produced by the *LBU_0363/LBU_1079*-encoded aspartate aminotransferase [4]. The expression of *LBU_1079* (average RPKM=1,500) was similar to the expression of the phosphoenolpyruvate carboxylase gene (*LBU_0412*) (**Table S2**) that catalyzed oxaloacetate formation, thus guaranteeing the de novo synthesis of aspartate and the subsequent synthesis of lysine, threonine, and asparagine in whey condition.

Lb. bulgaricus 2038 could synthesize cysteine from serine [3, 4]. However, the expression levels of all of the genes involved in serine synthesis were less than 100 RPKM (**Table S6**), indicating that de novo serine synthesis was not active under the whey condition. In addition, the average RPKM of *LBU_0247*, which encoded the enzyme catalyzing the last step of methionine synthesis, was only 16; , indicating that the methionine synthesis ability was weak. These evidences validated our postulation that under the dairy fermentation, this bacterial mainly imports amino acids from outside circumstance instead of de novo synthesis [5, 6].

Most efficient nitrogen source utilization system Coinciding with the degenerate amino acid synthesis pathway, *Lb. bulgaricus* 2038 possessed a rich protease and peptide/amino acid transport system. Clustering analysis of these protease and transporter system genes based on the different growth stages allowed us to separate them into four groups (**Figure 5, Table S7**). The largest group (Group 4) was highly expressed in the log and stationary stages, including two Opp systems, two ABC-type amino acid transport systems, PrtB, and 10 peptidases. This finding illustrated that strain growth during the late stage under whey growth conditions was mainly dependent on oligopeptides. In the control and lag stages, Group 1 included three highly expressed ABC-type amino acid/glutamine transport systems and Group 2 included a highly expressed ABC-type glutamine transport system. These results hint that the strain can sufficiently utilize free amino acids from the environment during the early growth stages. This finding is consistent with our previous finding that peptides and free amino acids in whey enabled rapid entry into the logarithmic growth phase and that the oligopeptide transport system was the primary pathway used to obtain amino acids [5].

Unique genes of Lb. bulgaricus 2038 Strain-specific genetic features of *Lb. bulgaricus* 2038 that facilitate its industrial application have been revealed. These genes include a unique lysine synthesis capability (*LBU_0931-LBU_0937*), the conversion of aspartate into carbon-skeleton intermediates (*LBU_0686*), formate production (*LBU_1742*), special exopolysaccharide (EPS) synthesis clusters (*LBU_1598-LBU_1588*, *LBU_1630-LBU_1618*) and complete type II restriction-modification (RM) systems (*LBU_0994-LBU_0995*) [3].

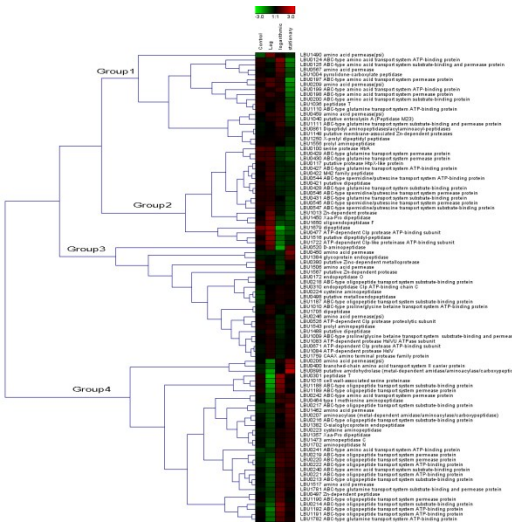


Figure 5. Hierarchical clustering of protease and transporter system genes. Log 2 transformed RPKM values of each gene were clustered by average linkage hierarchical clustering method. Red color represented high expression and green color represented low expression.

Molecular BioSystems

ARTICLE

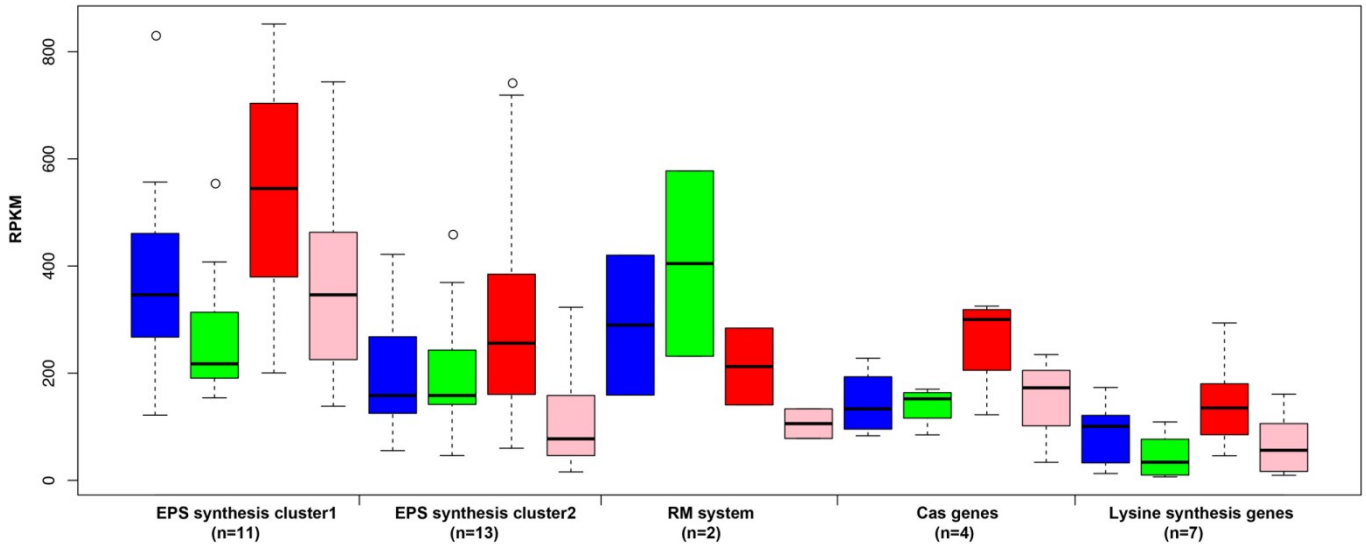


Figure 6 Transcriptome profile of functional categories composed of *Lb. bulgaricus* 2038 unique genes in four growing stages. Functional categories are listed in the x-axis with gene numbers in brackets. Numbers on the y-axis represents RPKM values. Blue, green, red, and pink represented control, lag, logarithmic, and stationary stages, respectively.

A total of 164 genes were revealed to be unique to *Lb. bulgaricus* 2038 relative to other *Lb. bulgaricus* strains (**Table S8**) [3]. Compared with the average level of gene expression (RPKM=723), these unique genes were expressed at a relatively low level (average RPKM=132). Among them, the genes of two EPS clusters were expressed at a medium level throughout the culturing process (**Figure 6**), which indicated that EPS was synthesized from the initiation of the culture. In addition to promoting texture, EPS could help the bacteria overcome gastrointestinal challenges and persist for longer periods in the gut [37, 38]. EPS from LAB have also been reported to have potential therapeutic applications, such as antitumor effects, immunostimulatory activity, and the ability to lower blood cholesterol [39]. Therefore, the persistent synthesis of EPS might contribute to the industrial application of *Lb. bulgaricus* 2038.

The other unique genes with medium expression levels included the type II RM system and the CRISPR-associated protein (Cas) genes (*LBU_0744-LBU_0747*). A 1,289 bp CRISPR loci (755,499-756,788) with 20 direct repeats (DRs) was identified downstream of the four Cas genes. The RM system and CRISPR/Cas system represent bacterial innate immunity and adaptive immunity, respectively [40], and the RM system had been reported to active against the virulent *Lb. delbrueckii* phage [41]. It was known that dairy lactobacilli could be inhibited by phages [42], and the attack of phages was common in the manufacture of yogurt [43]. So the persistent and stable low expression of these two types of defence systems in *Lb. bulgaricus* 2038 guaranteed its protection against phage infection and facilitated its use for industrial manufacturing.

The lysine synthesis genes showed low expression levels (average RPKM=85), coinciding with the conclusion that the main nitrogen usage pathway is through proteolysis instead of de novo synthesis.

The *Lb. bulgaricus* 2038 unique gene *LBU_0686* encoded the phosphoenolpyruvate carboxykinase gene that catalyzed the synthesis of phosphoenolpyruvate (PEP) from oxaloacetate (OAA) and endowed the strain with the ability to bring aspartate into the carbon cycle because LBU_0363 and LBU_1079 can convert aspartate into OAA [3, 4]. However, the expression of *LBU_0686* during all four stages remained very low (RPKM<20). This finding is similar to that observed with *LBU_0363*, hinting

that the aspartate into carbon-skeleton pathway is not functional in whey medium.

LBU_1742 encoded GTP cyclohydrolase II, which degrades GTP into formate. This gene was also maintained at a low expression level throughout the culture process (average RPKM=14), excluding this formate supply route. However, gene *LBU_1511* encoding the oxalate: formate antiporter was found to be highly expressed in whey, and ranked as the 5th most highly expressed gene during the stationary stage (**Table 2**). Thus, this pathway might supply the bacteria with formate through the exchange of oxalate with the environment.

Operon components and UTR regions

Operons are composed of consecutive genes on the same strand that are transcribed together and co-regulated. RNA-seq technology made operon prediction possible due to the availability of sequences transcribed from intergenic regions. We used Rockhopper [22] to determine the operon structures of *Lb. bulgaricus* 2038. This program takes into account the correlation of consecutive gene expression levels and intergenic distances. As a result, 372 operons composed of 1,125 genes (62.8%) were predicted throughout the *Lb. bulgaricus* 2038 genome (**Figure S4**, **Table S9**), with 200 operons on the positive strand and 172 on the negative strand. Thus, 81.5% of the operons (303) were located on the leading strand. A total of 88.4% of the operons were small transcriptional units comprising 2-4 genes (two genes- 55.6%, three genes- 22.3%, and four genes- 10.5%), while the largest operon was composed of 24 genes and the second two largest operons separately included 13 genes.

Genes in the same operon might execute the same functions[44], *e.g.*, ten proteins encoded by Operon_159 (*LBU_0761-LBU0772*) constituted the complete fatty acid biosynthesis pathway of *Lb.bulgaricus* 2038, and genes in several operons were significantly enriched in different papthways (**Table S10**). Co-transcription of genes involved in same pathway would save up energy and increase efficiency of both regulation and translation.

The 24 gene operon (Operon_072) encoded 21 ribosomal proteins and one translation initiation factor (IF-1), which were all involved in the COG Class of 'Translation, ribosomal structure and biogenesis'. All 24 of these genes were significantly up-regulated (FDR<0.001, fold change >2) during the logarithmic phase (average RPKM=12,146), and down-regulated during the stationary phase (average RPKM=2109), coinciding with the bacterial growth status (**Figure S5, Table S11**). The 1st 13 gene operon (Operon_135, *LBU_0628-LBU0640*) encoded five cell division proteins and five proteins involved in 'Cell envelope biogenesis'. The 2nd 13 gene operon (Operon_131, *LBU_0598-LBU0610*) encoded eight F0F1-ATPase system subunits that participate in oxidative phosphorylation in 'Energy production and conversion' (**Table S11**). The discrepant expression levels of the three operons were consistent with the bacterial physiological activity (i.e., the operon in charge of protein synthesis had the highest expression level (RPKM=5,778), the operon involved in energy production (RPKM=2,566) was expressed at a higher level, and the operon participating in cell division (RPKM=561) was expressed at a medium level).

A total of 547 transcription start sites (TSS) and 544 transcription termination sites (TTS) were also identified in this analysis (**Table S12**). The relative position of the TSS to the start codon was -54, and the average distance of the TTS to the stop codon was 60 bp. Because the cDNA synthesis protocol did not capture the 5' CAP structure of the mRNA, a part of the TSSs identified here may be derived from fragmented mRNA.

Helicase genes

We found 13 hypothetical protein genes that were co-expressed with the helicase genes (**Table S13**). Helicases unwind double-stranded DNA, self-annealed RNA, or RNA-DNA hybrids in all organisms [45] and are associated with fundamental biological functions such as ribosome assembly, translation initiation, and DNA replication. There were a total of 19 helicase genes in *Lb. bulgaricus* 2038, including three RNA helicase genes (**Table S14**). In our clustering analysis, all three RNA helicase genes were co-expressed with the aminoacyl tRNA synthetase genes, thereby validating the role of the RNA helicases in protein translation.

The auto-aggregation of *Lactobacillus reuteri* was reported to be mediated by the putative DEAD-box helicase AggH [46], which showed 66% identity with LBU_0288, an ATP-dependent RNA helicase in *Lb. bulgaricus* 2038. Therefore, we speculated that the high expression of helicase might also help *Lb. bulgaricus* 2038 aggregate and contribute a growth advantage. Based on our clustering results, we determined that several helicase genes were co-expressed with transporting genes (**Table S13**). *LBU_1514*, encoding a putative helicase, was co-expressed with the amino acid permease gene *LBU_0450*.

Another seven helicase genes (*LBU_0012*, *LBU_0025*, *LBU_0383*, *LBU_0707*, *LBU_1055*, *LBU_1170*, and *LBU_1175*) were all co-expressed with genes involved in ABC-type transport systems, such as the ABC-type oligopeptide transport system substrate-binding protein (**Table S13**). The function of both ABC transporters and helicases depended on the energy provided by ATP hydrolysis, and comparison of the 3D structures revealed that the folds of their nucleotide-binding sites are similar [47]. As the role of DNA helicase is to unwind duplex DNA during replication [48], the high expression of helicase genes might send a message to ABC transporters through competing ATP that the strain need nutrition to satisfy the growth need, thus helicase genes co-expressed with amino acids/oligopeptides transporting genes might mediate the ability of the strain to obtain nitrogen sources from the environment.

Additionally, seven helicase genes of *Lb. bulgaricus* 2038 (*LBU_0288*, *LBU_0452*, *LBU_0613*, LBU_0697, *LBU_0707*, *LBU_1175*, and *LBU_1369*) showed homology to human helicase genes, with protein identities ranging from 29%-38%. These genes included the three RNA helicases, and their conservation among different genomes reflected the important role of helicases in fundamental biological processes.

Pseudogenes

Lb. bulgaricus was reported to be involved in reductive evolution in nutritionally rich environments; this status was mainly reflected by the number of pseudogenes in the genome [35]. A total of 339 pseudogenes were identified in *Lb. bulgaricus* 2038 (**Table S15**), representing 18.9% of the predicted CDS. These pseudogenes were mainly truncated or split, including 79 fragments split from 37 genes. In our analysis, only 4.9% of the entire transcriptome was mapped to pseudogenes, indicated by a 4-fold reduction (4.9% versus 18.9%) in the expected transcriptional activity. In contrast, the pseudogene expression in *Salmonella typhi* showed a 10-fold reduction [10], suggesting that the reduction of pseudogenes in *Lb. bulgaricus* 2038 was not as severe as that reported in *Salmonella typhi*. Among these pseudogenes, only 31 exhibited a complete loss of transcriptional activity (RPKM<10) during fermentation (e.g., the three amino acid permease genes *LBU_0209*, *LBU_0459*, and *LBU_1490*). This result is in accordance with our previous conclusion that *Lb. bulgaricus* 2038 preferred the use of ABC type oligopeptide transport systems to amino acid permeases to obtain amino acids [4, 5]. The fact that 258 of the 339 (76.1%) pseudogenes had significant expression changes during the different growth stages suggested that the pseudogenes might also play roles in the living cell, perhaps through their remaining functional domains. This phenomenon might be caused by a recent degeneration in the coding region structure that did not affect the promoter region.

Among the split genes, 61 fragments were co-transcribed in 29 operons. This finding indicated that these split genes were still actively transcribed. A typical example was the lac repressor gene (*lacR*) in the lactose operon (*LBU_1027-LBU_1024*), which was split into two pseudogenes (*LBU_1025* and *LBU_1024*). Although the *lac* operon in *Lb. bulgaricus* was no longer regulated, the expression level of the two *lacR* pseudogenes was similar to *lacS* (*LBU_1027*) and *lacZ* (*LBU_1026*), suggesting that the inactivation of the *lac* repressor was achieved at the translational level.

We were not surprised to identify several pseudogenes that are transcribed at a high level (>1000 RPKM) in *Lb. bulgaricus* 2038 under whey growth conditions (**Table S15**). In *Mycobacterium leprae*, 25% of transcripts were derived from pseudogenes [49]; these pseudogenes have functional roles in infection, intracellular parasitization and replication [50]. Moreover, expressed pseudogenes have been proposed to function as a class of non-coding RNAs and act as riboregulators, thereby regulating gene expression at both the transcriptional and posttranscriptional levels [51].

Non-coding Transcriptome

Non-coding RNA in bacteria have been widely revealed to modulate a variety of physiological responses [52]. The strand-specific sequencing technology enabled us to study antisense RNA in *Lb. bulgaricus* 2038. We mapped all of the transcripts to 1,790 genes in the antisense orientation, and found that 0.69%-1.29% of transcripts were antisense in the four growth stages (**Table 1**).

Correspondingly, we found 103 genes with higher antisense RNA expression levels (antisense: sense ratio ≥1) in the four growth stages (**Table S16**) (e.g., *LBU_0042* involved in sucrose metabolism), hinting at the inhibitory role of non-coding RNAs in regulating gene expression. A total of 85 of these genes exhibited a high antisense RNA expression level during the control stage, whereas only 18, 2, and 9 genes exhibited high antisense RNA expression levels during the lag, logarithmic, and stationary stages, respectively. During the control stage, *Lb. bulgaricus*

2038 was transferred from glycerol stock cultures into whey medium and endured a sharp environmental change; thus, the majority of the highly expressed antisense RNAs were observed during this stage. Among the 103 genes regulated by antisense RNA, the most abundant functional class belonged to the transport system (25 genes, 24.3%), including 13 genes belonging to the ABC-type transport system, two amino acid permease genes, and nine other transporter genes. Another two large classes were the transposase genes (eight) and regulators (five).

We only identified one intergenic small RNA (sRNA) in the genome that was 210 nt in length (600864-600655). This sRNA was located 51bp downstream of *LBU_0613* and 91bp upstream of *LBU_0612*. sRNAs typically prevent translation and accelerate mRNA degradation by base-pairing with mRNA targets [53]. Target prediction indicated that this sRNA was likely to interact with *LBU_0667* (the minimum free energy of the sRNA-target interaction=-17.32 kJ/mol), which encoded the metallo beta subunit lactamase. This sRNA was expressed at a very high level (RPKM=20,520) during the control phase and declined sharply in the three subsequent growth phases (RPKM=654, 118, and 31, respectively). But the expression of *LBU_0667* (**Table S2**) only showed a slight increase from the control to the logarithmic stage, and the expression profile between sRNA and target gene in stationary phase did not show inverse correlation (**Figure S6**). This indicated that the target gene might has different regulation system upon entry into stationary phase, and most likely the sRNA played no important role in regulating mRNA stability since only one sRNA was identified while over 100 antisense RNAs were revealed in *Lb.bulgaricus* 2038. .

Because the food service industry does not allow the use of genetically modified bacteria, the study of the genetic and molecular mechanisms of LAB are hampered to a certain extent, and many barriers to genetic manipulation remain to be overcome [54]. Antisense RNA has the potential to reduce transcript and protein levels but not damage the targeted gene, thereby providing a method to modulate gene expression in LAB [55].

Conclusion

We applied RNA-seq technology to explore the transcriptome of *Lb. bulgaricus 2038*. Through our research, gene expression profiles at four time points were described and operons were predicted on a genome-wide scale. We revealed 372 operons composed of 62.8% of the genes in *Lb. bulgaricus* 2038, with genes constituting operons exhibiting similar expression profiles. The most abundantly expressed genes in the four stages were mainly involved in translation (logarithmic stage), glycolysis (control/lag stages), lactic acid production (all four stages) and 10-CHO-THF production (stationary stage). The expression of the unique RM system and CRISPR/Cas system implied the bacteriophage defense mechanism of *Lb. bulgaricus* 2038 and might contribute to its industrial application. Several industrial physiological features were explained by our transcriptome analysis, such as why D-lactic acid occupied 90% of the product of *Lb. bulgaricus* 2038, the important role of 10-CHO-THF in anabolism and purine biosynthesis, the highly expressed amino acid transport system during the early growth stage and the oligopeptide transport system during the late stage. In contrast, several other postulations (i.e., NADPH production through the GapN bypass, the conversion of aspartate into carbon-skeleton intermediates, and formate production through GTP degradation) were demonstrated to not be functional under whey growth conditions. The high expression of the helicase genes and the co-expression of the amino acid/oligopeptide transporting proteins might help *Lb. bulgaricus* 2038 aggregate and provide a growth advantage. Our study revealed the presence of sRNA and antisense RNA in *Lb. bulgaricus* 2038 and found that the most abundant genes regulated by antisense RNA belonged to the transport system. These results indicated that the strain has the potential to be regulated by non-coding RNA, thus providing a possible new rationale for the genetic manipulation of *Lb. bulgaricus*.

To the best of our knowledge, this is the first report to describe transcriptome data in *Lb. bulgaricus* using RNA-seq. The descriptions of the gene expression profiles in different culture stages, operon components, and non-coding RNAs will provide new insights into the industrial features of *Lb. bulgaricus* and improve our understanding of the genetic basis of LAB.

Acknowledgements

This work was sponsored by the Shanghai National Science Foundation (11ZR1425500). We thank Yongqiang Zhu and Yuezhu Wang of the Chinese National Human Genome Center at Shanghai for their kindly assistance with the RNA-seq work.

References

1. M. Sasaki, B. W. Bosman and P. S. Tan, *The Journal of dairy research*, 1995, **62**, 601-610.
2. E. S. D Beshkova, G Frengova, Z Simov., *Journal of Industrial Microbiology and Biotechnology*, 1998, **20**, 180-186.
3. P. Hao, H. Zheng, Y. Yu, G. Ding, W. Gu, S. Chen, Z. Yu, S. Ren, M. Oda, T. Konno, S. Wang, X. Li, Z. S. Ji and G. Zhao, *PloS one*, 2011, **6**, e15964.
4. H. Zheng, E. Liu, P. Hao, T. Konno, M. Oda and Z. S. Ji, *Biotechnology letters*, 2012, **34**, 1545-1551.
5. E. Liu, H. Zheng, P. Hao, T. Konno, Y. Yu, H. Kume, M. Oda and Z. S. Ji, *Current microbiology*, 2012, **65**, 742-751.
6. E. Liu, H. Zheng, P. Hao, T. Konno, H. Kume, L. Ye, M. Oda, K. Suzuki and Z. S. Ji, *International Dairy Journal*, 2014, **35**, 145-152.
7. S. Sieuwerts, D. Molenaar, S. A. van Hijum, M. Beerthuyzen, M. J. Stevens, P. W. Janssen, C. J. Ingham, F. A. de Bok, W. M. de Vos and J. E. van Hylckama Vlieg, *Applied and environmental microbiology*, 2010, **76**, 7775-7784.
8. Z. Wang, M. Gerstein and M. Snyder, *Nature reviews. Genetics*, 2009, **10**, 57-63.
9. J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens and Y. Gilad, *Genome research*, 2008, **18**, 1509-1517.
10. T. T. Perkins, R. A. Kingsley, M. C. Fookes, P. P. Gardner, K. D. James, L. Yu, S. A. Assefa, M. He, N. J. Croucher, D. J. Pickard, D. J. Maskell, J. Parkhill, J. Choudhary, N. R. Thomson and G. Dougan, *PLoS genetics*, 2009, **5**, e1000569.
11. C. M. Sharma, S. Hoffmann, F. Darfeuille, J. Reignier, S. Findeiss, A. Sittka, S. Chabas, K. Reiche, J. Hackermuller, R. Reinhardt, P. F. Stadler and J. Vogel, *Nature*, 2010, **464**, 250-255.
12. K. B. Arnvig, I. Comas, N. R. Thomson, J. Houghton, H. I. Boshoff, N. J. Croucher, G. Rose, T. T. Perkins, J. Parkhill, G. Dougan and D. B. Young, *PLoS pathogens*, 2011, **7**, e1002342.
13. I. Lasa, A. Toledo-Arana, A. Dobin, M. Villanueva, I. R. de los Mozos, M. Vergara-Irigaray, V. Segura, D. Fagegaltier, J. R. Penades, J. Valle, C. Solano and T. R. Gingeras, *Proceedings of the National Academy of Sciences of the United States of America*, 2011, **108**, 20172-20177.
14. M. M. Leimena, M. Wels, R. S. Bongers, E. J. Smid, E. G. Zoetendal and M. Kleerebezem, *Applied and environmental microbiology*, 2012, **78**, 4141-4148.
15. B. Lawley, I. M. Sims and G. W. Tannock, *Applied and environmental microbiology*, 2013, **79**, 5661-5669.
16. G. Giannoukos, D. M. Ciulla, K. Huang, B. J. Haas, J. Izard, J. Z. Levin, J. Livny, A. M. Earl, D. Gevers, D. V. Ward, C. Nusbaum, B. W. Birren and A. Gnirke, *Genome biology*, 2012, **13**, R23.
17. A. P. Vivancos, M. Guell, J. C. Dohm, L. Serrano and H. Himmelbauer, *Genome research*, 2010, **20**, 989-999.
18. P. Chomczynski and N. Sacchi, *Analytical biochemistry*, 1987, **162**, 156-159.
19. B. Langmead and S. L. Salzberg, *Nature methods*, 2012, **9**, 357-359.
20. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer and B. Wold, *Nature methods*, 2008, **5**, 621-628.
21. L. Wang, Z. Feng, X. Wang, X. Wang and X. Zhang, *Bioinformatics*, 2010, **26**, 136-138.
22. R. McClure, D. Balasubramanian, Y. Sun, M. Bobrovskyy, P. Sumby, C. A. Genco, C. K. Vanderpool and B. Tjaden, *Nucleic acids research*, 2013, **41**, e140.
23. A. Sturn, J. Quackenbush and Z. Trajanoski, *Bioinformatics*, 2002, **18**, 207-208.
24. F. Eggenhofer, H. Tafer, P. F. Stadler and I. L. Hofacker, *Nucleic acids research*, 2011, **39**, W149-154.
25. I. Grissa, G. Vergnaud and C. Pourcel, *Nucleic acids research*, 2007, **35**, W52-57.
26. S. Penaud, A. Fernandez, S. Boudebbouze, S. D. Ehrlich, E. Maguin and M. van de Guchte, *Applied and environmental microbiology*, 2006, **72**, 7445-7454.
27. K. J. Livak and T. D. Schmittgen, *Methods*, 2001, **25**, 402-408.
28. J. M. Toung, M. Morley, M. Li and V. G. Cheung, *Genome research*, 2011, **21**, 991-998.
29. S. Chandrasekaran and N. D. Price, *Proceedings of the National Academy of Sciences of the United States of America*, 2010, **107**, 17845-17850.
30. S. Karlin, J. Mrazek, A. Campbell and D. Kaiser, *Journal of bacteriology*, 2001, **183**, 5025-5040.
31. H. Willenbrock, C. Friis, A. S. Juncker and D. W. Ussery, *Genome biology*, 2006, **7**, R114.
32. P. M. Sharp and W. H. Li, *Nucleic acids research*, 1987, **15**, 1281-1295.
33. F. Coucheney, L. Gal, L. Beney, J. Lherminier, P. Gervais and J. Guzzo, *Biochimica et biophysica acta*, 2005, **1720**, 92-98.
34. A. Razeto, S. Kochhar, H. Hottinger, M. Dauter, K. S. Wilson and V. S. Lamzin, *Journal of molecular biology*, 2002, **318**, 109-119.
35. M. van de Guchte, S. Penaud, C. Grimaldi, V. Barbe, K. Bryson, P. Nicolas, C. Robert, S. Oztas, S. Mangenot, A. Couloux, V. Loux, R. Dervyn, R. Bossy, A. Bolotin, J. M. Batto, T. Walunas, J. F. Gibrat, P. Bessieres, J. Weissenbach, S. D. Ehrlich and E. Maguin, *Proceedings of the National Academy of Sciences of the United States of America*, 2006, **103**, 9274-9279.
36. Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa and M. Kanehisa, *Nucleic acids research*, 2007, **35**, W182-185.
37. S. Fanning, L. J. Hall, M. Cronin, A. Zomer, J. MacSharry, D. Goulding, M. O. Motherway, F. Shanahan, K. Nally, G. Dougan and D. van Sinderen, *Proceedings of the National Academy of Sciences of the United States of America*, 2012, **109**, 2108-2113.
38. E. Denou, R. D. Pridmore, B. Berger, J. M. Panoff, F. Arigoni and H. Brussow, *Journal of bacteriology*, 2008, **190**, 3161-3168.
39. K. V. M. a. K. V. PRABHAKAR, *ORIENTAL JOURNAL OF CHEMISTRY*, 2014, **30**, 1401-1410.
40. S. T. Abedon, *Bacteriophage*, 2012, **2**, 50-54.
41. L. Auad, M. A. A. Peril, A. Holgado and R. R. Raya, *Current microbiology*, 1998, **36**, 271-273.
42. A. O. Kilic, S. I. Pavlova, W. G. Ma and L. Tao, *Applied and environmental microbiology*, 1996, **62**, 2111-2116.
43. P. Munsch-Alatossava and T. Alatossava, *Frontiers in microbiology*, 2013, **4**, 408.
44. H. Salgado, G. Moreno-Hagelsieb, T. F. Smith and J. Collado-Vides, *Proceedings of the National Academy of Sciences of the United States of America*, 2000, **97**, 6652-6657.

45. S. S. Patel and I. Donmez, *The Journal of biological chemistry*, 2006, **281**, 18265-18268.
46. S. Roos, S. Lindgren and H. Jonsson, *Molecular microbiology*, 1999, **32**, 427-436.
47. C. Geourjon, C. Orelle, E. Steinfels, C. Blanchet, G. Deleage, A. Di Pietro and J. M. Jault, *Trends in biochemical sciences*, 2001, **26**, 539-544.
48. T. M. Lohman, *Molecular microbiology*, 1992, **6**, 5-14.
49. T. Akama, K. Suzuki, K. Tanigawa, A. Kawashima, H. Wu, N. Nakata, Y. Osana, Y. Sakakibara and N. Ishii, *Journal of bacteriology*, 2009, **191**, 3321-3327.
50. K. Suzuki, N. Nakata, P. D. Bang, N. Ishii and M. Makino, *FEMS microbiology letters*, 2006, **259**, 208-214.
51. V. A. Erdmann, M. Z. Barciszewska, A. Hochberg, N. de Groot and J. Barciszewski, *Cellular and molecular life sciences : CMLS*, 2001, **58**, 960-977.
52. L. S. Waters and G. Storz, *Cell*, 2009, **136**, 615-628.
53. D. Balasubramanian and C. K. Vanderpool, *RNA biology*, 2013, **10**, 337-341.
54. M. J. Kullen and T. R. Klaenhammer, *Current issues in molecular biology*, 2000, **2**, 41-50.
55. K. Bouazzaoui and G. LaPointe, *Journal of microbiological methods*, 2006, **65**, 216-225.