농 학 박 사 학 위 논 문

Multi-genomics approach to evolutionary and functional characteristics of lactic acid bacteria

유산균의 진화적 및 기능적 특성에 대한 다중 유전체학적 접근

2022 년 8 월

서울대학교 대학원
농생명공학부

전 수 민

유산균의 진화적 및 기능적 특성에 대한 다중유전체학적 접근

지도교수 김 희 발

이 논문을 농학박사 학위논문으로 제출함

2022 년 8 월

서울대학교 대학원
농업생명과학대학 농생명공학전공

전 수 민

전수민의 농학박사 학위 논문을 인준함
2022 년 8 월

위 원 장	김 영 훈	(Seal)
부위원장	김 희 발	(Seal)
위 원	박 태 섭	(Seal)
위 원	조 서 애	(Seal)
위 원	유 재 웅	(Seal)

Abstract

Multi-genomics approach to evolutionary and functional characteristics of lactic acid bacteria

Soomin Jeon

Department of Agricultural Biotechnology

Seoul National University

In this study, genome comparison analysis and microbial experiments were performed to obtain a comprehensive understanding of the bacterial evolution and their functionality at the molecular level. Lactic acid bacteria (LAB) are a species-rich in useful functions for humans and are widely used as health functional foods. Since LAB have many useful functions for humans, it is valuable to study. In addition, it is suitable for genome research as its short genome make it easier to understand the entire genome compared to other individuals. Therefore, research on the genome of LAB will not only increase the utilization of resources useful to humans, but also contribute to the genetic understanding of higher organisms with complex genomes. Therefore, this study was conducted to provide a multifaceted understanding of the evolution and functionality of LAB useful in humans through various genome analyses.

In Chapter 2, the full-length genome sequence of *Lactiplantibacillus plantarum* GB-LP3 (*Lactobacillus plantarum*), a functional lactic acid bacterium, was sequenced and its evolutionary characteristics were detected by comparing the published *L. plantarum* complete genomes. It was confirmed that it has the closest evolutionary distance to the genome of *L. plantarum* ZJ316 identified in infant fecal samples, and possesses several functional genes and an evolutionarily accelerated ATP transporter. Based on comparative analysis, it was possible to infer the adaptation of *L. plantarum* in the special environment such as Korean fermented food.

In Chapter 3, it was confirmed that *Lactobacillus delbrueckii* subsp. *bulgaricus* and *Limosilactobacillus fermentum*, which are widely used as health functional foods, have higher GC content compared to their genome size compared to other LAB species. It was confirmed that the high GC content was due to the difference in the 3rd nucleotide of the triplet code encoding the amino acid, and the difference in energy caused by this was compared. Through this, it was inferred that *L. bulgaricus* and *Lm. fermentum* evolved toward having a high GC content to adapt to the environment.

In Chapter 4, in order to confirm the phenotypic and genotype changes for the evolutionary pressure, a *Lactobacillus acidophilus* strain with increased heat resistance was developed by artificially exposing it to a high temperature environment. The heat adapted strain showed a significant

increase in survival rate at a high temperature of 65 degrees or more compared to the wild-type. Two SNPs were found in the vicinity of genes related to the cell wall through whole-genome comparison. Based on these results, it was suggested that the *L. acidophilus* strain evolved in the direction to harden the cell wall to adapt to heat stimulation.

In Chapter 5, the cognitive ability of mice fed *Lactobacillus acidophilus*, *Lacticaseibacillus paracasei*, and *Lacticaseibacillus rhamnosus* for 8 weeks was evaluated and changes in intestinal microbial composition were compared to confirm the effect of LAB on the experimental animal *in vivo*. Among the experimental groups, the group fed *L. acidophilus* showed the highest cognitive ability improvement, and 16 bacterial species showed a significant difference in the intestinal microbial flora comparison comparing control group. Many of the bacteria with the changed ratio are involved in the production of substances for the synthesis of neuronal substances acting on the animal's brain. Based on these results, evidences that orally ingested LAB can affect cognitive ability were indicated.

From Chapters 2 to 5 of this dissertation, the evolution and functional characteristics of LAB were presented through 3rd generation sequencing and genomic analysis. In detail, *de novo* whole-genome sequencing, phylogenetic tree construction, genome comparison analysis, and metagenome analysis were performed and these were applied to

understanding for LAB. These studies will provide a deeper understanding of the evolution and characteristics of microorganisms through sequencing.

Through these studies, it was possible not only to present the functional and evolutionary characteristics of LAB through genome analysis but also to identify the expected functionalities through experiments and to infer genetic factors. Through this research, a comprehensive understanding of the characteristics of microorganisms and genome analysis was provided.

Keyword : Genome sequencing, 3rd generation sequencing, Microbial evolution, Lactic acid bacteria, *Lactobacillus*

Student Number : 2016-21743

Contents

List of Tables

List of Figures

Chapter 1. Literature Review

1.1. Bioinformatics

1.1.1. Genome sequencing

Genome sequencing is the process of decoding DNA to A, T, G, and C in order. Since DNA is the most basic unit of living organisms, genome sequencing is a necessary process to fundamentally understand living organisms. Genome sequencing was performed for the first time in history by the Maxam-Gilbert method and the Sanger method. The Maxam-Gilbert method uses base-specific chemical degradation as the most primitive method of deciphering a base sequence (Maxam and Gilbert 1977). The 5'-phosphate of single-stranded DNA is labeled with radioactive ^{32}P and treated with four different chemicals that cleavage four types of nucleotide sequences. Fragments of different lengths are generated through the cleavage, and the nucleotide sequence behind the truncated nucleotide sequence is identified by electrophoresis. The Sanger method uses the DNA synthesis mechanism developed by Frederick Sanger and uses 2,3-dideoxy nucleoside triphosphate (ddNTP) (Sanger, Nicklen et al. 1977). Unlike normal deoxynucleotides (dNTP), ddNTPs do not have a hydroxyl group at the 3'-end to elongate DNA synthesis. When the DNA synthesis reaction is induced using 4 test tubes containing each ddNTP, DNA sequences of different sizes are synthesized. These DNA fragments are sorted through electrophoresis, and the DNA sequence is decoded by this process. After being introduced in 1977, it was applied to an automated sequencer through

technological advances and contributed greatly to the human genome project (HGP). With the advent of the sequencing method, biological and medical research and discovery have been greatly accelerated. However, the initial sequencing method had limitations in that the cost was enormous and the producible data was small at once. Although the accuracy is high relatively, it is difficult to decipher the DNA of the whole organism using this method that can decipher only short sequences.

This difficulty has been solved by a novel sequencing method known as Next Generation Sequencing (NGS or Second-generation sequencing) (Grada and Weinbrecht 2013). NGS refers to massive parallel sequencing that which vast amounts of genomic information can be quickly deciphered by splitting a genome into many fragments, reading each fragment at the same time, and then assembling them (Zhou, Ren et al. 2010). NGS sequencing process usually consists of three steps: sample preparation-clonal amplification-sequencing reaction. Representative NGS methods include pyrosequencing and Illumina sequencing using Reversible Dye Terminator. Pyrosequencing is a method of deciphering the base sequence by detecting pyrophosphate produced when DNA polymerase synthesizes one nucleotide (Ahmadian, Ehn et al. 2006). During DNA synthesis, pyrophosphate is released when dNTP binding to the template enters. Adenosine-5'-phosphosulfate (APS) is converted to ATP by ATP sulfurylase of the pyrophosphate, and the generated ATP emits light by luciferase to

identify the base sequence. Among the commercial sequencer, Roshe 454 and GS FLX system uses this method. A sequencing platform using this method was commercialized in 1999, and the era of NGS began in earnest. Illumina sequencing is the most representative sequencing method in NGS (Chen, Dong et al. 2013). Similar to Sanger sequencing, it uses terminator molecules in which the hydroxyl group at 3' of the ribose is blocked. At the start of sequencing, a DNA template is combined with a primer having a sequence complementary to an adapter, and a polymerase is bound to the double-stranded DNA. A mixture of 4 fluorescent-labeled dNTPs is added at every cycle and each dNTP binds to an elongating complementary strand and fluoresces. dNTP identification is achieved through total internal reflection fluorescence microscopy with two or four laser channels. Illumina's platform is representative: Miseq is mainly used for amplicon sequencing, while Hiseq with a long production read length is mainly used for whole-genome sequencing. The data produced by the NGS method have different characteristics from data obtained through the previous methods (Goodwin, McPherson et al. 2016). Data produced by the NGS method is large-capacity data. Since the analysis complexity increases as the size of the increased data, a high-speed analysis algorithm is required. At the same time, since the length of the sequence is shorter than the data produced through Sanger sequencing, an assembly algorithm is required. The data of Sanger sequencing is expressed as the sum of all cell DNA, whereas in NGS

data, the nucleotide sequence of single-stranded DNA derived from each cell is expressed independently. Therefore, NGS data has issues with polymerase error. To overcome this problem, many researchers have tried to increase sequencing coverage (or depth), which indicates how many times a base has been read least. However, since the read length is short, the problem remains that the genome cannot be completely assembled including many repeat sequences or gaps.

The 3rd generation sequencing method is a novel NGS technology, which can sequence without PCR amplification (Karst, Ziels et al. 2021). Compared to NGS, it has a longer read length and can be analyzed with only a very small amount of DNA. Unlike NGS, since there is no PCR amplification process, the time and cost occurring in the PCR amplification process can be reduced, and errors occurring in the process can be prevented. Pacific Biosciences (PacBio) RSII and Oxford Nanopore MinION are representative third-generation platforms. PacBio RSII reads the nucleotide sequence as it is in a single molecule (Rhoads, Au et al. 2015). This is the technology called a single molecule, real-time (SMRT). It is a method of deciphering the base sequence by fixing DNA polymerase and detecting that the DNA passes through the polymerase. When a base added fluorescence passes, the fluorescence falls off and emits light. Because no terminator is required, sequencing can be performed at a faster speed than NGS, and errors from polymerase can be reduced since it does not undergo DNA

amplification. It produces a long read on average 1,000 base pairs. Although the error issue from DNA polymerase is removed, the error from the process of deciphering the fluorescent signal into the nucleotide sequence is maintained. Whereas Oxford Nanopore's MinION uses a nano-sized enzyme as a nanopore reader to decipher the base sequence by measuring the potential difference through which DNA molecules pass (Lu, Giordano et al. 2016). As DNA passes through the membrane where the current is generated by the movement of ions, the flow of the current is interrupted. Since the potential difference generated for each base is different, the potential difference at this time is measured. Compared to NGS, the read length is longer and the speed is faster. Also, unlike other equipment, its simple body is the advantage. However, due to the limitations of systemic error, homopolymer and repeat sequence decoding are weak. In order to overcome the error rate of 3rd generation equipment, a hybrid genome sequencing method using a 3rd generation sequencer and other platforms together is sometimes used. The second-generation platform has relatively high accuracy but produces short leads, so it is combined with the 3rd generation platform with a low accuracy but long leads to take advantage of each. It is possible to produce results with similar accuracy at a lower price than using a single platform. A mixture of PacBio and Illumina or a mixture of Nanopore and Illumina is often used.

1.1.2. Genomic data analysis

Due to the development of sequencing technology and cost reduction, it is now popular enough that individuals can decipher and analyze the genome sequence. The reads generated after decoding the nucleotide sequence are combined into a contig through the assembly process. The assembly program is chosen suitably depending on the sequencing method. Data produced by PacBio is usually assembled using programs such as HGAP and Canu, and one produced by Nanopore often uses Canu, Flye, and raven (Chin, Alexander et al. 2013, Koren, Walenz et al. 2017, Kolmogorov, Yuan et al. 2019, Vaser and Šikić 2021). The assembled contig is evaluated as N50, which means the length of the scaffold, which is the middle priority when the assembled scaffolds are arranged in order of length. The contig created through the assembly process goes through the polishing process using software such as quiver, Nanopolish, Racon, and Pilon (Walker, Abeel et al. 2014, Firtina, Kim et al. 2020, Hu, Huang et al. 2021). Previously, it was referred to as attaching the Illumina output to the generated contig, which produced high-accuracy but shorter reads, but now it refers to the case of attaching raw reads to the contig generated by PacBio or Nanopore to modify it. Through this process, the gap and uncertainty in the genome sequence are compensated. The generated genome is finally ready for analysis. The gene content included in the genome is checked through the annotation process. Since it searches a gene compared to the established

database, the quality of the database affects the annotation results. Databases such as DAVID (Functional Annotation Bioinformatics Microarray Analysis) for human genes and Swiss-Prot are mainly used for microbial genes (Bairoch and Boeckmann 1991, Dennis, Sherman et al. 2003). Through this process, it is possible to understand the basic genetic characteristics of an individual. However, genes not included in the database cannot be identified and structural features cannot be found. With the development of sequencing technology, a large amount of sequencing data has been accumulated, and methods for analyzing them have also been diversified. Indeed, it became possible to compare and analyze multiple genomes instead of analyzing a single genome. Comparative genome analysis is a method that compares multiple genomes to discover the genomic characteristics of a specific individual. It is mainly classified into an orthology analysis based on sequence homology at the gene level and synteny analysis at the genome level. According to analysis purposes, it is divided into evolutionary analysis, individual specificity discovery analysis, and metagenomic analysis. A representative analysis method used to understand the evolution of multiple individuals is the construction of a phylogenetic tree (Kapli, Yang et al. 2020). A phylogenetic tree is a method of inferring evolutionary history by representing a tree-shaped figure based on similarities and differences between individuals. In a phylogenetic tree, a pattern of branches spreading from one trunk can be seen, where the trunk

can be interpreted as a common ancestor and the ends of numerous branches can be interpreted as an evolved species or group. For phylogenetic analysis, after selecting the taxa to be compared, an outgroup is set to classify the taxa. Setting the outgroup makes it clearer how the taxon derives from its original ancestor. In general, the outgroup uses the sister group of the analyzing taxon. After that, the sequences are aligned and an evolution model is selected. Algorithms used to build phylogenetic trees include neighbor-joining, maximum likelihood, and Bayesian inference methods. There are two types of data used for creating a phylogenetic tree: the whole genome and specific genes. In the case of using the whole genome sequence, it is used to compare approximate similarity among the individuals by the average nucleotide identity (ANI), aligned WGS, and orthologous gene set (Yoon, Ha et al. 2017). In the other case, a clear aim of analysis exists. To distinguish between species, a phylogenetic tree is created by comparing 16S (or 18S) rRNA genes. To compare the evolution rate of a specific gene, a phylogenetic tree of the gene is created. dN/dS analysis is mainly used as a method for estimating the rate of evolution of a protein. It is a method of inferring the evolutionary selection of proteins by calculating the ratio of the rate at which nonsynonymous mutation occurs (dN) and the rate at which synonymous mutation occurs (dS) (Kryazhimskiy and Plotkin 2008). If there are more nonsynonymous mutations that do not change amino acids even when single nucleotide mutations occur, it can be interpreted as having

undergone negative selection. It acts to preserve the original state and maintain the function of the protein sequence of the orthologous protein. On the other hand, if a single nucleotide mutation causes more frequent synonymous mutations causing an amino acid change, it can be interpreted that the protein has undergone positive selection. This paralogous gene acts to acquire a new function or immunity. Since dN/dS analysis is recommended to be used in a section with high sequence conservation, analysis is mainly performed after orthologous gene definition. Pangenome (Pan-genome; supra-genome) refers to a set of different genes possessed by all organisms belonging to a single phylogenetic taxon (Vernikos, Medini et al. 2015). In microbiology, it is usually discussed limitedly at the species level, and it is analyzed by classifying it into a core gene that exists in all lineages and a variable gene group (accessory genes) that exists only in some lineages of a species. If genes increase as increasing the number of sequences (individuals), it is defined as an open pangenome. In particular, species living in diverse environments of mixed microbial populations continue to expand the pangenome because they have different ways of exchanging genetic material. Conversely, if the addition of the individual genome no longer provides new genes to the pangenome, it is defined as a closed pangenome. It usually occurs in species that live in isolated areas with limited access to microorganisms.

Method for detecting the genetic specificity of an individual includes SNP detection, functional gene categorization, and gene cluster detection. SNP detection is a method to find the difference between a single nucleotide sequence among aligned sequences. It is usually searched after orthologue gene definition and alignment and mainly uses a tool such as SnpEff and SAMtools (Li, Handsaker et al. 2009, Cingolani, Platts et al. 2012). Functional gene categorization is a method for comparing the number of functional genes. After sequencing, annotated gene groups are classified into functional gene groups by comparing them to databases such as COG, SEED, KEGG, Pfam, and GO. In the case of microorganisms, many tools that classify after annotation and provide a figure or table exist as WGS inputs. However, there are differences in classification methods and results depending on the database. Gene cluster detection is a method used to select a group of genes with a specific function. Existing well-known gene groups such as Bacteriocin and CRISPR can be easily searched for because their gene sequences have already been established. Entering the whole genome sequence provides results in a short time into a tool that is usually provided as a web-based tool. Frequently used tools include BAGEL to search for bacteriocin, CRISPRFinder to search for CRISPR genes, VRprofile to search for virus factors, and Islandviewer to search for genomic islands (de Jong, van Hijum et al. 2006, Grissa, Vergnaud et al. 2007, Bertelli, Laird et

al. 2017, Li, Tai et al. 2018). Using these analyzed tools, the basic genetic composition contained in the genome can be identified.

1.2. Lactic acid bacteria

1.2.1. Lactic acid bacteria as probiotics

Lactic acid bacteria (LAB) are bacteria that produce lactic acid as a fermentation product by consuming carbohydrates such as glucose and are gram-positive, acid-tolerant, and nonsporulating. The bacteria have the property of inhibiting the growth of harmful bacteria by lactic acid, and they also have beneficial effects on the host's body when they inhabit the intestines of mammals. Therefore, LAB are called probiotics, which are living microorganisms that are beneficial to health when consumed in an appropriate dose. Since LAB are used as probiotics live in fermented foods or animal gastrointestinal tract, they basically have properties such as acid, bile, and salt resistance. *Lactobacillus* is a representative LAB that has been used as probiotics and is a gram-positive strain that produces lactic acid as a result of fermentation (Duar, Lin et al. 2017). Its existence and efficacy were first suggested by the Russian scientist Ilya Mechnikov. It is mainly found in fermented foods such as yogurt and cheese, and endemic species are often found in kimchi and doenjang, which are traditional Korean foods (Lee, Yoon et al. 2011, Jung, Jung et al. 2016). Because it mainly inhabits an environment with high osmotic pressure and strong acid, its basic characteristics are high acid resistance and high osmotic pressure. Well-

studied strains include *Lactobacillus rhamnosus* GG, *Lactobacillus acidophilus* La-14, and *Lactobacillus plantarum* 299v. Previously, the genus *Lactobacillus* was a huge group containing more than 250 species, but the *Lactobacillaceae* family, which previously contained three genera, was changed to a giant family containing 25 genera using genomic information in 2020 (Zheng, Wittouck et al. 2020). Due to this reclassification, *Lacticaseibacillus casei* and *Limosilactobacillus reuteri*, which were previously commonly used as *Lactobacillus*, belong to a new family. Bifidobacterium, which is also classified as LAB, previously belonged to *Lactobacillus* but was isolated (Felis and Dellaglio 2007). Similarly, it seems that research on LAB has been accumulated and subdivided.

1.2.2. Functionality and role of LAB

The positive effects of LAB as probiotics on human health are well known (Sanders and Klaenhammer 2001, De Angelis, Calasso et al. 2016). When oral ingested LAB affect the intestinal flora, it prevents the growth of harmful bacteria in the intestine and regulates immunity to normal levels. Besides, it is indicated that LAB improve autoimmune skin diseases such as atopic dermatitis, suppress hypertension and metabolic syndrome, and lower insulin resistance. Its positive effects on the host's health are due to the basic characteristics of LAB, which secrets lactic acid (Tachedjian, Aldunate et al. 2017). Due to lactic acid, the pH of the environment containing the LAB-

containing community is lowered. This has the effect of inhibiting other harmful bacteria. In addition, short-fatty chain acids (SCFAs) such as butyrate produced by LAB in the intestine and substances such as GABA are known to affect the host body (Komatsuzaki, Shima et al. 2005, Mao, Li et al. 2019). Recently, the Brain-Gut axis and Gut-Lung axis theories have been proposed that changes in the composition of intestinal microbes affect the host body such as the brain and lungs, and experimental research results are also reported (Dumas, Bernard et al. 2018, Liu, Liong et al. 2018). Interest in the effects of LAB is expected to increase in the future.

1.2.3. NGS for LAB application

Since the genome of LAB is very small (2MB-3Mb) and it is a monoploid, it is very easy to decipher the genome. Although the microbial genome is smaller in size than the human genome, it contains all the genetic information necessary for survival. Decoding the genome of microorganisms will not only understand the basic genetic mechanism of living things but will also help understand the human genome. In addition, since the characteristics of microorganisms vary according to the environmental niche and lifestyle, complete decoding of the microbial genome is important academically and industrially. Currently, decoding the genome of a single microorganism with technologies such as PacBio and Nanopore is very inexpensive and can be performed quickly. Furthermore, it

is possible to decode the genome of a mixture of the flora of which species are not identified. By identifying the mixed genomes contained in a sample, it can analyze the microbial community of a specific environment, which is called metagenomics. 16S rRNA (or 18S rRNA) is usually used for identification. 16S rRNA gene of the mixed genomes is amplified by PCR, and the amplified sequence is identified by comparing the sequence of a known microorganism. Indeed, the metagenome analysis through genome analysis is very useful because microbial communities are often impossible to separate and identify experimentally due to their non-culturable. This research method is widely used in the ratio of intestinal microbes and related clinical studies. The previously reported correlation between obesity and the intestinal microflora was also revealed through metagenomic analysis of the gut microbiome. The relationships with host allergy, diabetes, and autoimmune diseases were also suggested by metagenomic analysis of the gut microbiome.

The genetic range of microorganisms is vast. Microorganisms are the group that has evolved the longest on Earth, and it is estimated that the number of bacteria that have been identified so far is about 1% of all microorganisms on Earth. At the same time, microorganisms are exposed to extreme environments and have the characteristics of rapid mutation. Therefore, deciphering and analysis of the microbial genome will not only become the basic foundation for understanding living organisms at the

genetic level but will also greatly contribute to understanding their evolutionary aspects.

This chapter was published in *Infection, Genetics and Evolution* (2017)
as a partial fulfillment of Soomin Jeon's Ph.D program.

Chapter 2. Comparative genome analysis of *Lactobacillus plantarum* GB-LP3 provides candidates of survival-related genetic factors

2.1. Abstract

Lactobacillus plantarum is found in various environmental niches such as in the gastrointestinal tract of an animal host or a fermented food. This species isolated from a certain environment is known to possess a variety of properties according to inhabited environment's adaptation. However, a causal relationship of a genetic factor and phenotype affected by a specific environment has not been systematically comprehended. *L. plantarum* GB-LP3 strain was isolated from Korean traditional fermented vegetable and the whole genome of GB-LP3 was sequenced. Comparative genome analysis of GB-LP3, with other 14 *L. plantarum* strains, was conducted. In addition, genomic island regions were investigated. The assembled whole GB-LP3 genome contained a single circular chromosome of 3,206,111bp with the GC content of 44.7%. In the phylogenetic tree analysis, GB-LP3 was in the closest distance from ZJ316. The genomes of GB-LP3 and ZJ316 have the high level of synteny. Functional genes that are related to prophage, bacteriocin, and quorum sensing were found through comparative genomic analysis with ZJ316 and investigation of genomic islands. dN/dS analysis identified that the gene coding for phosphonate ABC transporter ATP-binding protein is evolutionarily accelerated in GB-LP3. Our study found that potential candidate genes that are affected by environmental adaptation in Korea traditional fermented vegetable.

Keywords

Bacterial survival, Comparative genomics, dN/dS, Genomic islands, *Lactobacillus plantarum*

2.2. Introduction

Lactobacillus plantarum, facultative anaerobic lactic acid bacteria, mainly inhabits various environmental niches including the gastrointestinal tract of an animal host or a fermented food (Johansson, Molin et al. 1993, Sanni, Morlon-Guyot et al. 2002). This species was often found in extreme condition including low pH and high osmotic pressure environments (McDonald, Fleming et al. 1990, Glaasker, Tjan et al. 1998). Studies have indicated that *L.plantarum* isolated from a specific environment often possesses a variety of properties according to adaptation for an inhabited environment. A study demonstrated that *L. plantarum* C88 strain isolated from traditional Chinese fermented dairy tofu has a potential antioxidant (Li, Zhao et al. 2012). Other study indicated that *L. plantarum* C-11 strain isolated from a cucumber fermentation was shown to be bacteriocidal (Daeschel, McKenney et al. 1990). Kwak et al. indicated that lactic acid bacteria from Kimchi including *L. plantarum* has anticancer activity (Kwak, Cho et al. 2014).

In order to search a ground of function in a specific environment, genomic analysis based on biological knowledge is needed. As the function of a living organism is biologically a result of a final occurrence according to the central dogma, the fundamental reason could lie in the genome (Crick 1970, Shapiro 2009). It is known from previous studies that there is a connection between environment and genes (Khoury and Wacholder 2009,

Beaty, Ruczinski et al. 2011). Many studies have been published in order to identify the strain-specific factor of lactic acid bacteria at the genomic level, since several strains of *L.plantarum* are found in various and extreme environments although same species. Kleerebezem et al. indicated that several genes provide an important part of the interaction with its environment by sequencing and characterization of *L. plantarum* WCFS1 strain genome (Kleerebezem, Boekhorst et al. 2003). Chaillou et al. insisted that genes related to biofilm formation may assist its adaptation to a specific environment by identifying a genome of *Lactobacillus sakei* 23K (Chaillou, Champomier-Vergès et al. 2005). However, a causal relationship of genetic factor and phenotype have not been systematically comprehended according to the specific environment.

In order to cumulate data to explain environmental adaptation at the genomic level, we investigated a new strain of *L.plantarum* GB-LP3, which was isolated from Korean traditional fermented vegetable. To understand the genomic relationship with the specific environment at the molecular level, this study aimed to identify *L.plantarum*'s environmental specific genetic predisposition by comparative analysis, dN/dS analysis, and genomic islands analysis. This study may provide a deeper insight to understand the interaction of genetic factor and phenotype according to a specific environment's adaptation.

2.3. Materials and Methods

2.3.1. Sample isolation and whole genome sequencing

GB-LP3 was isolated from Korean traditional fermented vegetable. Genomic DNA was isolated and purified using an UltraClean Microbial DNA Isolation Kit (MoBio,Carlsbad, CA, USA). The concentration and purity were measured by NanoDrop spectrophotometer (Thermo Scientific, Wilmington,DE, USA). Approximately 5 μg of extracted genomic DNA was cut into 8–12 kb fragments using a Hydroshear system (Digilab, Marlborough, MA, USA). SMRTbell libraries were prepared for SMRT (Single Molecule RealTime) sequencing with C4 chemistry on a PacBio RS II system (Pacific Biosciences, Menlo Park, CA, USA). Libraries were purified using 0.45× AMPure XP beads to eliminate short inserts of <1.5 kb. The size distribution of the sheared DNA template was characterized using an Agilent 12000 DNA Kit (Applied Biosystems, Santa Clara, CA, USA), and concentration was determined using Invitrogen Qubit (Carlsbad, CA, USA). Primers of the sequencing were annealed to the templates at a final concentration of 5 nM template DNA, and DNA polymerase enzyme C4 was added according to the manufacturer's instructions for small-scale libraries. A DNA/Polymerase Binding Kit P6 (Pacific Biosciences) was used to load the enzyme template-complexes and libraries onto 75,000 zero-mode waveguides. A DNA Sequencing Reagent 2.0 Kit (Pacific Biosciences) was used to sequence SMRT cells using a 120-min sequence capture protocol

along with a stage start to maximize the subread length with PacBio RS II. Raw sequence data from the PacBio RS II system were filtered and assembled using the PacBio SMRT portal system ver. 2.3.0. The "RS_HGAP_assembly.3" algorithm was employed and the genome size parameter was set to 3,300,000 bp using the Compute Minimum Seed Read Length option whereas other parameters were set to default. Assembled contigs with a short contig length (<20,000 bp) and low coverage (<50×) were filtered for further analysis. To remove errors in the pre-assembled GB-LP3 genome sequence, an iterative polishing process was conducted until no genomic variants were identified.

2.3.2. Annotation and identification of GB-LP3 genome

GB-LP3 genome was annotated using RAST server with default parameters (Aziz, Bartels et al. 2008). The GB-LP3 genome was visualized by DNAplotter (Carver, Thomson et al. 2009). For functional annotation, protein coding sequences were predicted and were categorized in SEED subsystem using RAST. Additional COG annotation was carried out using COGNIZER (Bose, Haque et al. 2015). To investigate GB-LP3 specific genes, genomic islands were found using IslandPath included in IslandViewer (Langille and Brinkman 2009).

2.3.3. Comparative genome analysis

To compare GB-LP3 with same species, the genome sequences of 24 *L.plantarum* strains were downloaded from the NCBI database. Among the sequences, genomes of 10 strains were at chromosome level (FMNP10, DSM 13273, PS128, ATCC 14917, JCM 1149, L31-1, UCMA 3037, NL42, SF2A35B, and DSM 16365). The genomes of other 14 strains were at complete genome level (ST-Ⅲ, ZJ95, JBE245, 5-2, B21, ZS2058, WCFS1, JDM1, DOMLa, P-8, 16, ZJ316, Zhang-LL, and HFC8). Average Nucleotide Identity (ANI) values of each pair were calculated using JSpecies (Richter and Rosselló-Móra 2009). In order to demonstrate the evolutionary distance between *L.plantarum* strains, ANI trees were built by MEGA7 (Kumar, Stecher et al. 2016). Maximum likelihood method was used for constructing trees.

To investigate GB-LP3 specific genes, orthologous genes were found. Sequences of 15 strains were aligned using MESTORTHO and PRANK (Kim, Sung et al. 2008, Löytynoja and Goldman 2008). To remove poorly aligned sequences, Gblocks was used (Castresana 2000). Genes containing stop codons within the coding sequence were removed. A tree of orthologous genes was constructed using MEGA7. The orthologous gene tree was constructed by neighbor-joining method with a bootstrap value of 1,000. To present how similar two genomes, SyMAP was used to show the similarity between genomes (Soderlund, Bomhoff et al. 2011). To search genes contained in only GB-LP3 compared to ZJ316, two genomes was

compared using IslandPick included in IslandViewer, which can identify unique regions by comparing a user-specified genome against closely related genomes. Unmatched regions in GB-LP3 with ZJ316 were detected using BLAST and regions less than 10,000bp were filtered.

2.3.4. dN/dS Analysis

In order to find evolutionary accelerated genes of GB-LP3, dN/dS value of each orthologous gene sets was calculated using PAML (Yang 2007). We calculated probability under alternative hypothesis and null hypothesis of branch-site model using codeml. Mean ω ratio (dN/dS ratio) was calculated by multiplying each probability by each predicted ω value. Using two log likelihood values obtained from codeml, likelihood ratio test was performed and P values were corrected by BH method. In this process, sequences with mean ω value under 1.0 were eliminated. Also, sequences that the difference of two log likelihoods is negative and that adjusted P value is over 0.05 were removed.

The three-dimension structure of a protein was predicted using Cn3D, and a location of amino acid was detected by counting the sequence (Hogue 1997). Information of ABC protein ArtP in complex with AMP-PNP/Mg^{2+} was used and this information of 3D structure was downloaded from RCSB PDB website. Its PDB ID is 3C41.

2.4. Results

2.4.1. General features of *L. plantarum* GB-LP3

The circular chromosome of *L. plantarum* GB-LP3 is comprised of 3,206,111 bp in size, 44.7% of GC content, 16 rRNA and 72 tRNA. Fig. 1 illustrates the GB-LP3 genome including tRNA and rRNA. Genome annotation at the RAST server shows that GB-LP3 genome encodes 3153 protein coding sequence.

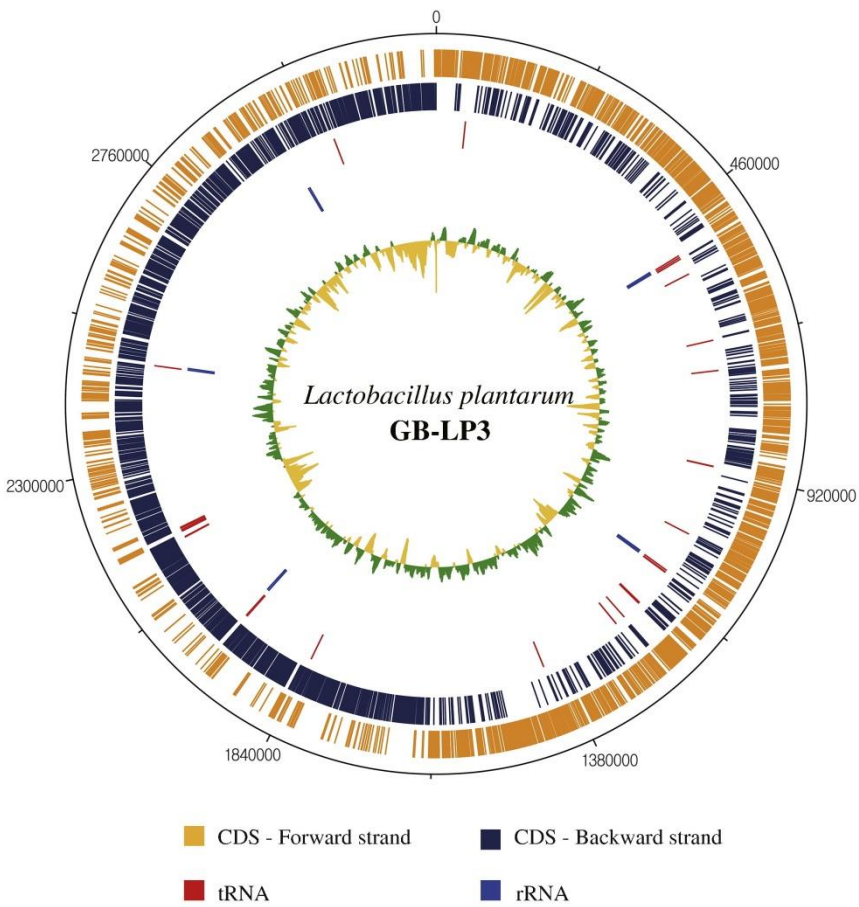


Figure 2-1. The genome map of *Lactobacillus plantarum* GB-LP3.

The outer five circles are CDSs of forward strand (light orange color), CDSs of a backward strand (dark blue color), tRNA (orange color), rRNA (blue color), GC contents (green and yellow color), respectively.

The functional categorization of the annotated protein by SEED subsystem and COG annotation is shown in Fig. 2. In the SEED subsystem, 2237 genes (70.95%) encode known functional proteins whereas 916 genes (29.05%) encode hypothetical proteins. The largest proportion of protein coding categories in SEED subsystem are “Carbohydrates” (388 genes) and “amino acids and derivatives” (235 genes). In the categorization of COG annotation, 2337 genes (74.12% of total protein coding sequences) were classified into COG annotation categories. The largest proportion of protein coding sequence is “general function prediction only” (562 genes) followed by “Amino acid transport and metabolism” (447 genes). The smallest portion is occupied by “cell motility” which has 8 genes. Both of circular graphs represent the categorization of predicted protein coding sequences in GB-LP3 genome. They show (A) SEED categorization using RAST and (B) COG annotation.

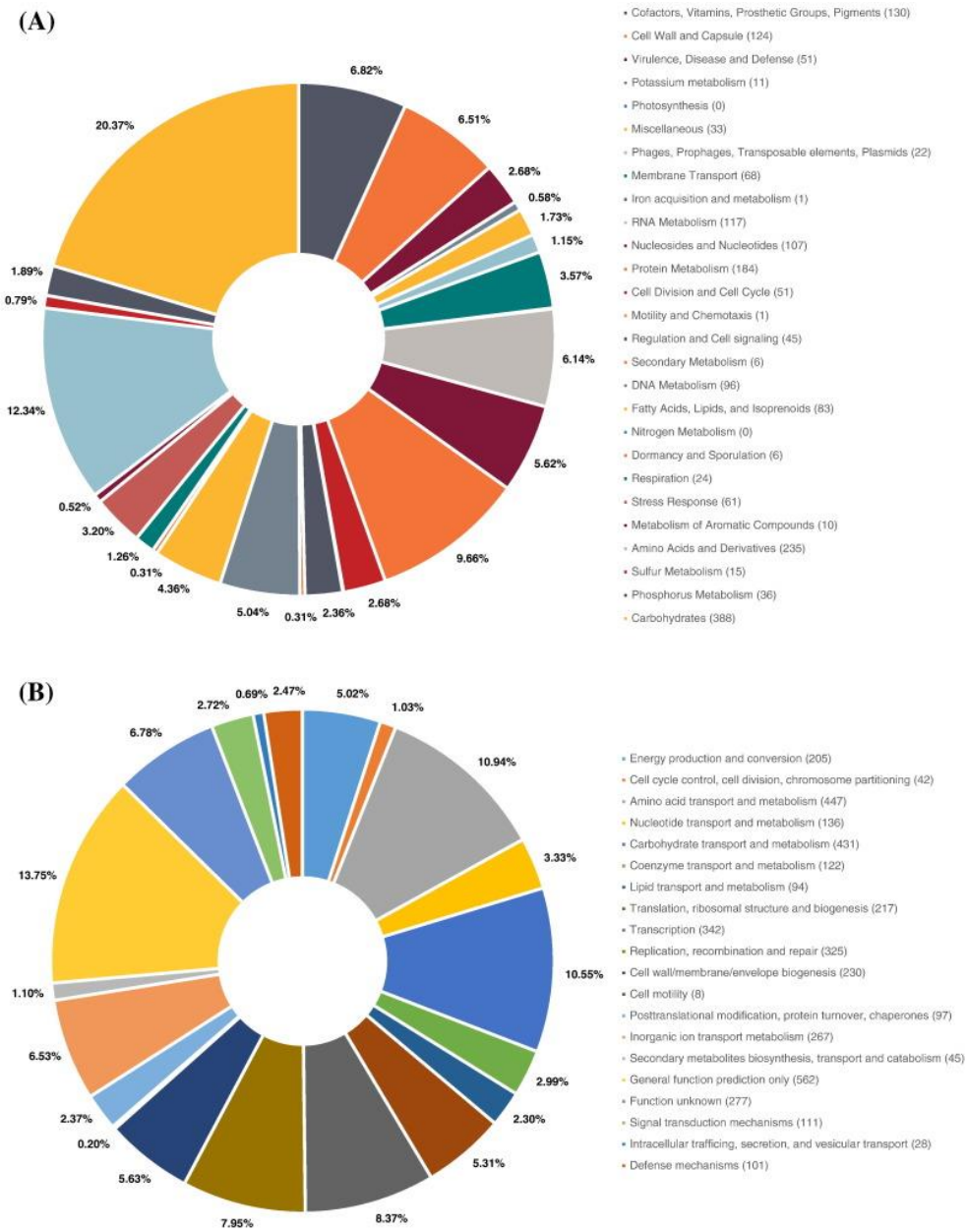


Figure 2-2. Functional categorization of all predicted protein coding sequences in the strain GB-LP3.

Both of circular graphs represent the categorization of predicted protein coding sequences in GB-LP3 genome. They show (A) SEED categorization using RAST and (B) COG annotation.

2.4.2. Phylogenetic trees among *L. plantarum* strains

Three phylogenetic trees were constructed for the comparative analysis of GB-LP3 within species. Two phylogenetic trees were constructed using ANI value. The tree in Fig. 3A was generated for 10 genomes of the chromosome level and 15 complete genomes. L31-1, ZJ316, UCMA 3037, P-8, and 16 were clustered with GB-LP3. Among the strains, ZJ316 was found to be the closest sequence to GB-LP3. SF2A35B and DSM 16365 presented the farthest evolutionary distance from GB-LP3 and shared the least similarity with GB-LP3. Fig. 3B was generated for 15 complete genomes. In this tree, P-8, 16, and ZJ316 strains were in the same clade with GB-LP3. It shows that the closest strain to GB-LP3 is ZJ316 (ANI value is 99.51) and the farthest strain is HFC8 (ANI value is 98.71).

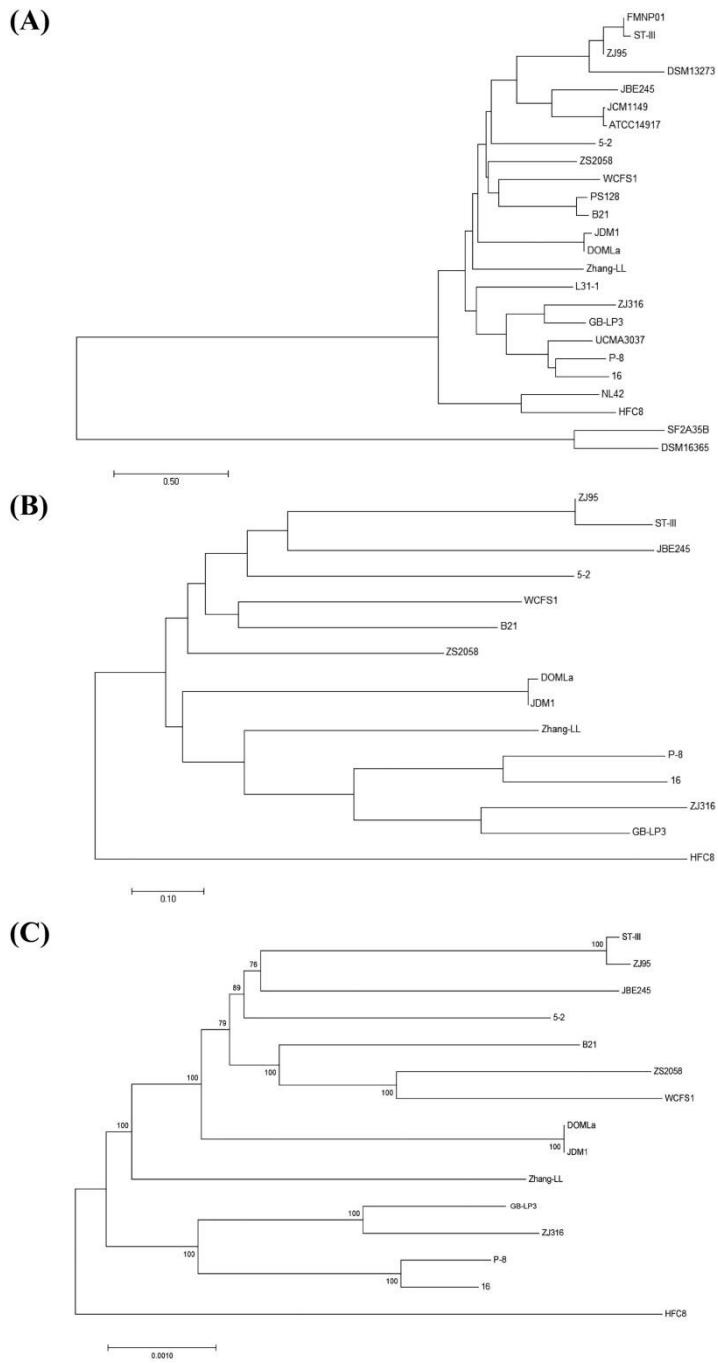


Figure 2-3. Phylogenetic trees between *Lactobacillus plantarum*.

(A) The ANI tree of 10 available genome sequences and 15 available complete genome sequences. **(B)** The ANI tree of 15 available complete genome sequences. **(C)** The orthologous gene tree of 15 available complete genome sequences.

Phylogenetic tree was constructed using 1848 orthologous genes of 15 complete genomes (Fig. 3C). It shows similar aspects to Fig. 3A and B in that GB-LP3 was clustered with ZJ316, P-8, and 16. GB-LP3 shows the closest evolutionary distance from ZJ316 but the farthest distance from HFC8. Even though few clustering patterns were not identical in several strains, the general appearance of evolutionary relation in the two ANI and orthologous gene trees (Fig. B and C) are similar. Therefore, the general structure of orthologous gene tree is reliable.

2.4.3. Comparative genome analysis with *L. plantarum* ZJ316

Comparative analysis was performed between GB-LP3 and ZJ316 which was found to be the closest strain to GB-LP3 according to phylogenetic tree analysis. The general features of the two genomes are presented in Table 1. Even though not completely linear, there is high level of synteny between the two genomes as can be seen in Fig. 4. The posterior region of GB-LP3 genome sequence of approximately 2,950,000 bp matched with anterior region of the ZJ316 genome.

Table 2-1. Genomic information of GB-LP3 and ZJ316

Genome	*GB-LP3*	*ZJ316*
Genome size (bp)	3,206,111	3,299,755
GC contents (%)	44.7	44.4
Number of subsystems	333	333
Number of coding sequences	3,153	3,213
Number of tRNA	72	61
Number of rRNA	16	15

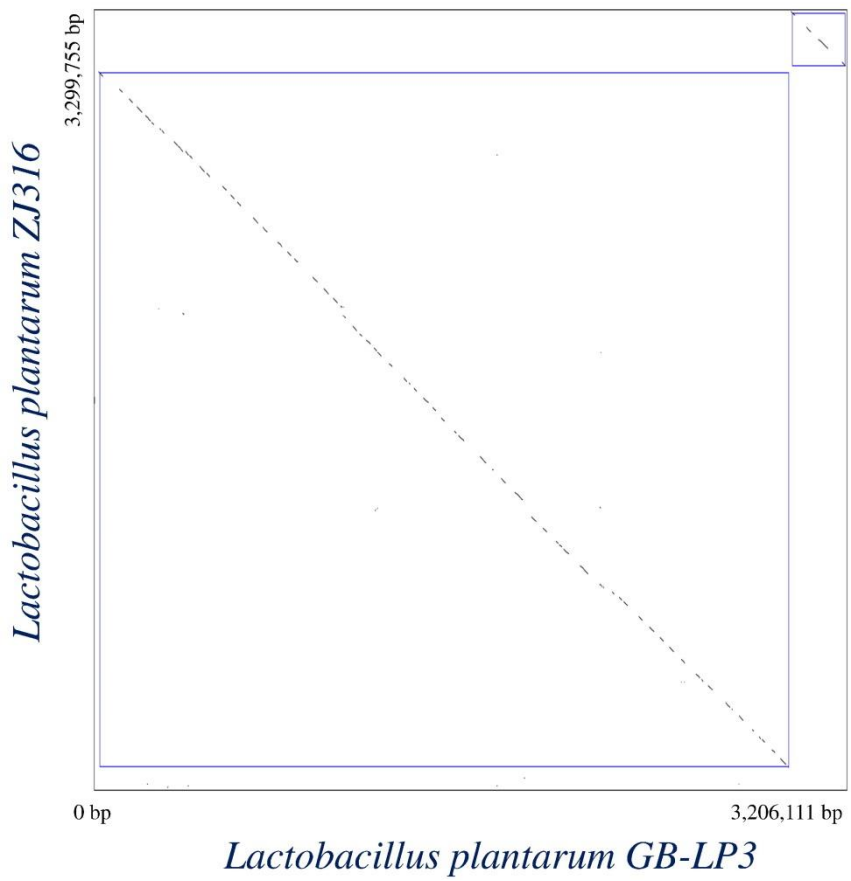


Figure 2-4. Dot plot for comparison between GB-LP3 and ZJ316 genomes

Horizontal axis is GB-LP3 genome and vertical axis is ZJ316 genome. It used only single chromosome except plasmids.

Five islandpick regions which are identified as unique regions by comparing GB-LP3 against the ZJ316 genome are found only in GB-LP3 were investigated as shown in Fig. 5 as green lines. The length of the largest region is 19,559 bp. Several genes related to phage and prophage formation were located in these regions including integrase, SaPI, prophage Lp2, Lp3, Lp4, and Lj965 protein, phage terminase, phage protein, phage-like repressor, phage tail length tape-measure protein, phage tail fibers, phage capsid and scaffold protein, phage integrase (site-specific recombinase), hypothetical SAV0792 homolog in SaPI, and hypothetical protein. Additionally, 12 unmatched regions in GB-LP3 compared with ZJ316 over 10,000 bp were detected and they contain 237 genes including prophage genes. These regions are shown in Fig. 5 as sky-blue colored lines.

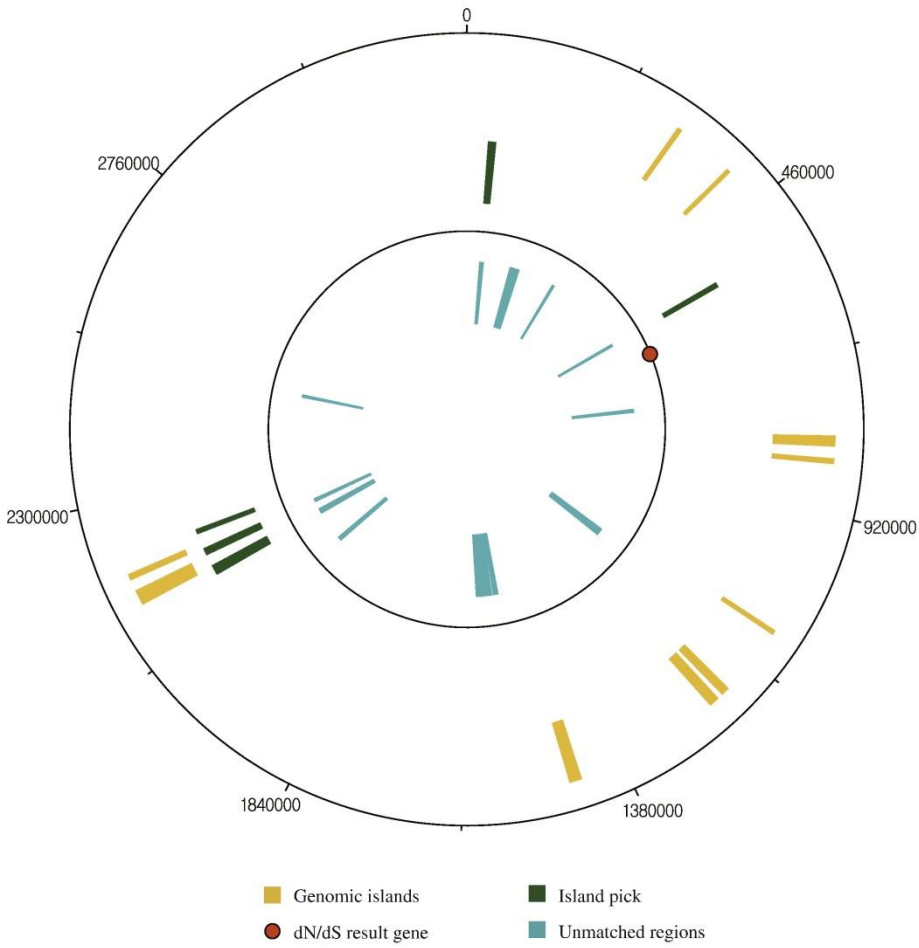


Figure 2-5. GB-LP3 strain-specific genes

GB-LP3 strain-specific genes include genomic island regions, island pick (unique regions in GB-LP3), and a gene of dN/dS result. The three lines show genomic islands (yellow color), island pick (green color), and unmatched regions (sky-blue color), respectively. The orange colored circle represents the dN/dS result gene. Island pick represents unique gene regions which are only GB-LP3 genome not in ZJ316 as the result of analysis using IslandPick.

2.4.4. Investigation of GB-LP3 specific genes

To identify evolutionarily accelerated genes in GB-LP3, dN/dS analysis was conducted using orthologous genes. Among the 1848 orthologue genes, one gene is identified as evolutionary accelerated gene as dN/dS value (mean ω) is 2.747255476 (< 0.05) as listed Table 2 and Fig. 5 as a result of the branch-site model. This result is regarded reliable because P value which provides how significant a result is 0.012083446. This gene encodes phosphonate ABC transporter ATP-binding protein (ABC transporter alkylphosphonate). The total length of phosphonate ABC transporter ATP-binding protein gene of GB-LP3 was 768 nucleotides (256 amino acids). In GB-LP3 genome sequence, one nucleotide variant site was found in only GB-LP3 comparing with other *L. plantarum* 14 strains. Among the 14 variants found by comparing sequences of 15 strains, the 665th nucleotide variant was found in GB-LP3 sequence (Table 3). The 665th nucleotide of the sequence is adenine while that of others is guanine (Fig. 6A). As a result, the 222nd amino acid of this gene of GB-LP3 is an aspartic acid which is negatively charged and acidic, but it is glycine in other genome sequences. As it is presented in Fig. 6B, the 222nd amino acid is located in a α-helix positioned outside of protein in a three-dimensional structure.

Table 1-2 Information of evolutionary accelerated gene (phosphonate ABC transporter ATP-binding protein)

Category	*Value*
Subsystem	Phosphonate metabolism
ω2a of foreground	565.88927
ω2a of background	0.25848
Proportion of 2a	0.00440
Mean ω	2.747255476
P value	0.012083446
FDR value	0.012083446

Table 2-2. All variants of the dN/dS gene among 15 strain sequences

Strain	*94*	*166*	*283*	*287*	*303*	*312*	*330*	*579*	*665*	*666*	*674*	*683*	*717*	*728*
5-2	G	A	A	C	G	C	G	G	G	G	T	G	G	G
Zhang-LL	T	A	A	C	G	C	G	G	G	G	T	G	G	G
DOMLa	G	G	A	T	G	C	G	G	G	G	T	G	G	G
ZS2058	G	A	G	C	A	C	G	G	G	G	T	G	G	G
16	G	A	A	C	G	A	A	G	G	C	C	G	A	G
GB-LP3	G	A	A	C	G	C	G	G	**A**	C	T	G	A	G
ST- III	G	A	A	C	G	C	G	G	G	C	T	G	A	G
P-8	G	A	A	C	G	A	A	G	G	C	C	G	A	G
ZJ95	G	A	A	C	G	C	G	G	G	C	T	G	A	G
ZJ316	G	A	A	C	G	C	A	G	G	G	T	G	G	G
B21	G	A	A	C	G	C	G	G	G	G	T	G	G	G
JBE245	G	A	A	C	G	C	A	G	G	G	T	A	G	G
HFC8	G	A	A	C	G	C	A	G	G	G	T	G	G	A
WCFS1	G	A	A	C	G	C	A	A	G	G	T	G	G	G
JDM1	G	G	A	T	G	C	G	G	G	G	T	G	G	G

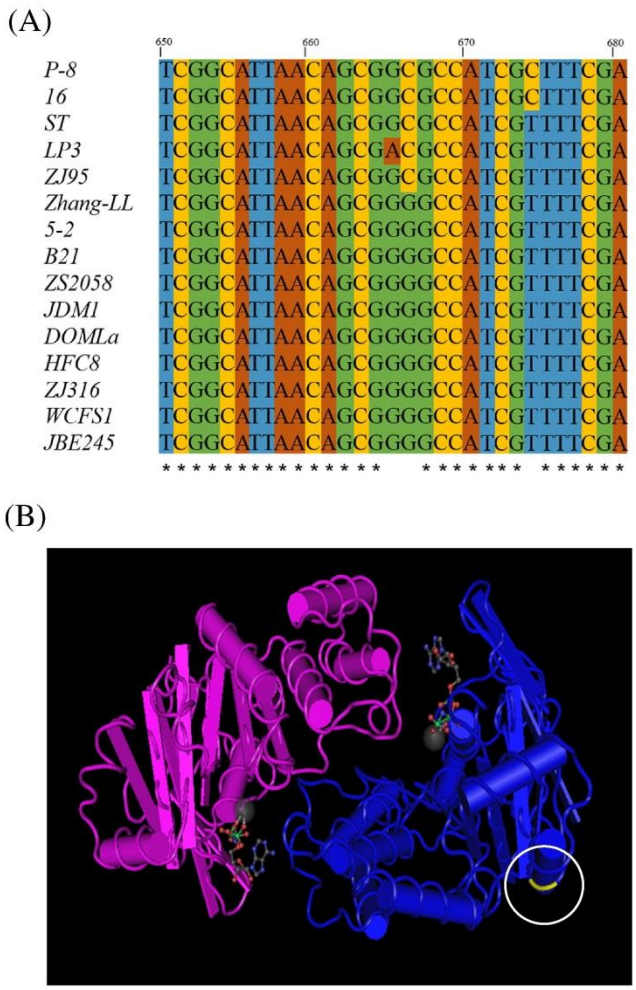


Figure 2-6. Variation of phosphonate ABC transporter ATP-binding protein

(A) Variant of nucleotides in phosphonate ABC transporter ATP-binding protein genes. **(B)** The figure shows three dimensional structure of the protein. The site of yellow color has different amino acid in GB-LP3 comparing *L. plantarum* strains.

Investigation of genomic islands (GEIs) of GB-LP3 was conducted and 10 GEIs were found. GEIs are parts of a genome that have an evidence of horizontal gene transfer, which frequently contains virulence genes (Dobrindt, Hochhut et al. 2004). These regions are shown in Figure 5 as yellow colored lines. Factors which have the ability to transfer genes were found, but no virulence factor in GB-LP3 genome. Most of the genes were classified into six types including those associated with phage formation, DNA expression, quorum sensing, integrase/recombinase, oligopeptide, and bacteriocin. The number of genes related to phage formation is the highest.

2.5. Discussion

2.5.1. Genomic islands in GB-LP3 genome

GEIs are formed as a result of horizontal gene transfer, which are known to code functions including pathogenesis, symbiosis, and adaptation of microbial. GEIs in bacteria genome are known as they carry genes that offer selective advantages to the host (Dobrindt, Hochhut et al. 2004). For example case of beneficial GEIs, there was a previous study which investigated many bacteria that have obtained genomic diversity through the acquisition of GEIs. As a result, they found out that GEIs have been advantageous to the survival of the bacteria, leading to bacterial evolution (Juhas, van der Meer et al. 2009). In this study, 2 of GEIs included functional genes which are related to bacteriocin and quorum sensing, which are expected that they were beneficial to the GB-LP3 in two ways for its function. Bacteriocin-related genes are expected to confer antimicrobial activity to GB-LP3 which may help this bacteria win their competition over other bacteria in the fermented environment. On the other hand, quorum sensing related genes are predicted to help regulate the density of bacterial population by secreting certain molecules and affect bacterial survival and evolution (Kuipers, de Ruyter et al. 1998, Miller and Bassler 2001, Sturme, Francke et al. 2007). Considering these results, we speculated that GEIs in GB-LP3 are advantageous to its survival.

2.5.2. Genetic factors related to survival fitness

Prophage genes including Lp1, Lp2, Lp3, and Lp4 were found in the GB-LP3 genome from comparative genomic analysis and GEI investigation. A previous study suggested that prophage genes provide a selective advantage to the lysogenic host through prophage-host interaction by increasing the immunity and superinfection against phage infection (Desiere, McShan et al. 2001). Here, Desiere et al. explained that mutated prophage genes in the bacterial genomes, such as *Streptococcus pyogenes* strain SF370, induce Darwinian coevolution of prophages and lysogenic bacteria. Another study experimentally showed that prophage genes were postulated to contribute to increasing the survival fitness of the lysogenic clone (Ventura, Canchaya et al. 2003). It showed that Lp1, Lp2, R-Lp3, and R-Lp4 genes were transcribed in *L.plantarum* WCFS1 and some of these - Lp2, R-Lp3, and R-Lp4 - are defective prophages. Although specific mechanism and role of prophage genes have not been studied, these studies are an indirect evidence that prophage genes might function to increase the host survival ability, when the host adapting to a specific environment. Therefore, we suggest that prophage genes in GB-LP3 genome could provide advantages and help to increase its survival in fermented food.

The dN/dS analysis revealed phosphonate ABC transporter ATP-binding protein gene as the evolutionarily accelerated gene. This protein is known for having organic phosphonate transmembrane transporter activity

and taking part in the phosphonate pathway and C-P lyase system (Huang, Su et al. 2005). According to the branch-site model, the variant site was located on the surface of the protein. The surface of transporter protein often interacts with plasma membrane and other regulatory factors related to post-modification (Higgins 1992, Hicke and Dunn 2003). Thus, the change of amino acid residue might affect the interaction of this protein with plasma membrane or regulatory factors. Besides, phosphonate (substrate of this protein) is an essential element of living organisms, as it constitutes nucleotides, phospholipids, and is a source of phosphorus (Karl 2000). To utilize phosphonate, C-P lyase complex is needed together with the phosphonate ABC transporter (Hirota, Kuroda et al. 2010). Because the concentration of phosphate is low in fermented environment where GB-LP3 inhabits, we suggest that evolutionarily accelerated gene in GB-LP3 could affect is survival in the presence of other microbes to its advantages and guide adaptation in fermentation environment (Sanchez and Demain 2002, Hsieh and Wanner 2010, Ren, Feng et al. 2013, Hove-Jensen, Zechel et al. 2014).

In this study, GEI investigation and comparative genomic analysis using the complete sequence of GB-LP3, which is a new *L.plantarum* strain, indicate that prophage genes and the phosphonate ABC transporter ATP-binding protein gene might affect the survival of GB-LP3. Although these factors are not a reason that GB-LP3 can live in fermented food comparing

with other microbes living in similar environments, all GB-LP3 specific genes are related to its survival. It indicates that these genes are candidates of genetic factors for survival. Therefore, we suggest that these survival-related factors could function as helpful keys to give advantages and help to adapt in fermentation environment. This study will provide deeper insight into understanding environmental adaptation of *L.plantarum* at genomic level.

Additional file 1 – Table S2-1. Genetic regions in only *L. plantarum* GB-LP3

Genomic islands region	Gene start	Gene end	Gene size	Strand	Product
1	37623	38777	1155	-1	Intergrase, superantigen-encoding pathogenicity islands SaPI
1	40326	41126	801	1	Prophage Lp4 protein7, DNA replication
1	41126	42520	1395	1	Prophage Lp3 protein 8, helicase
1	42667	43086	420	1	Prophage Lp4 protein 12
1	43110	43421	312	1	Phage terminase large subunit
1	43408	43749	342	1	Prophage Lp3 protein 11
1	43859	44131	273	1	phage protein
1	45032	45505	474	1	Phage terminase small subunit
1	45502	47205	1704	1	prophage Lp3 protein 15, terminase large subunit
1	47360	48466	1107	1	phage portal protein
1	48456	50036	1581	1	prophage Lp3 protein 18
1	50151	50420	270	1	prophage Lp3 protein 19, head-to-tail joining
1	50579	50935	357	1	prophage Lp3 protein 20
2	528752	529897	1146	-1	Intergrase
2	531201	531884	684	1	prophage Lp4 protein 3, phage-like repressor
2	231938	532204	300267	1	prophage Lp4 protein 4
2	532516	532938	423	1	prophage Lp4 protein 5
2	532931	533152	222	1	prophage Lp4 protein 6

2	533202	533945	744	1	prophage Lp4 protein 7, DNA replication
2	533957	535375	1419	1	Prophage Lp3 protein 8, helicase
2	535986	536360	375	1	prophage Lp4 protein 11, DNA replication
2	536362	536775	414	1	prophage Lp4 protein 12
3	2130309	2136620	6312	-1	prophage Lp2 protein 53
3	2136635	2136997	363	-1	Lj965 prophage protein
3	2137011	2142842	5832	-1	prophage Lp4 protein 6phage tail length tape-measure protein
3	2142858	2143061	204	-1	Lj965 prophage protein
3	2143229	2143627	399	-1	Lj965 prophage protein
3	2143727	2144197	471	-1	phage tail fibers
3	2144212	2144577	366	-1	Lj965 prophage protein
3	2144577	2145128	552	-1	Lj965 prophage protein
3	2145130	2145477	348	-1	Lj965 prophage protein
3	2148778	2150463	1686	-1	phage capsid and scaffold
3	2150836	2152344	1509	-1	phage portal protein
3	2152356	2153594	1239	-1	phage terminase, large subunit
3	2153584	2154462	879	-1	Phage terminase small subunit
4	2174366	2174737	372	-1	prophage Lp3 protein 20
4	2174895	2175164	270	-1	prophage Lp3 protein 19, head-to-tail joining
4	2175574	2177115	1542	-1	prophage Lp3 protein 18
4	2177122	2178212	1091	-1	phage portal protein
4	2178367	2180070	1704	-1	prophage Lp3 protein 15, terminase large subunit

4	2180067	2180540	474	-1	Phage terminase small subunit
4	2181236	2181508	273	-1	phage protein
4	2181618	2181956	339	-1	phage protein
4	2181943	2182134	192	-1	prophage Lp3 protein 10
4	2182149	2182628	480	-1	prophage Lp3 protein 9
4	2182774	2184168	1395	-1	Prophage Lp3 protein 8, helicase
4	2184168	2184968	801	-1	prophage Lp4 protein 7, DNA replication
4	2185328	2185471	144	-1	prophage Lp3 protein 5
4	2185479	2185658	180	-1	prophage Lp3 protein 4
4	2186589	2187344	756	1	prophage Lp3 protein 1, integrase
4	2187367	2187747	381	1	phage integrase: site-specific recombinase
5	2218369	2218728	360	-1	hypotehtical SAV0792 homolog in superantigen-encoding pathogenicity islands SaPI
5	2219025	2220656	1632	-1	DNA primase/helicase, phage-associated
5	2221446	2221727	282	-1	prophage Lp4 protiein 4
5	2222922	2224088	1167	1	integrase, superantigen-encoding pathogenicity island SaPI

Additional file 2 – Table S2-2. Genomic island pick of *L. plantarum* GB-LP3

Locus	Type	Start	End	Strand	Frame	Product
chromosome	CDS	37623	38777	-	0	Integrase%2C superantigen-encoding pathogenicity islands SaPI
chromosome	CDS	38856	39500	-	0	hypothetical protein
chromosome	CDS	39649	39828	+	1	hypothetical protein
chromosome	CDS	40111	40329	+	1	hypothetical protein
chromosome	CDS	40326	41126	+	0	prophage Lp4 protein 7%2C DNA replication
chromosome	CDS	41126	42520	+	2	prophage Lp3 protein 8%2C helicase
chromosome	CDS	42667	43086	+	1	prophage Lp4 protein 12
chromosome	CDS	43110	43421	+	0	Phage terminase%2C large subunit
chromosome	CDS	43408	43749	+	1	prophage Lp3 protein 11
chromosome	CDS	43859	44131	+	2	Phage protein
chromosome	CDS	45032	45505	+	2	Phage terminase small subunit
chromosome	CDS	45502	47205	+	1	prophage Lp3 protein 15%2C terminase large subunit
chromosome	CDS	47159	47359	+	2	hypothetical protein
chromosome	CDS	47360	48466	+	2	Phage portal protein
chromosome	CDS	48456	50036	+	0	prophage Lp3 protein 18
chromosome	CDS	50151	50420	+	0	prophage Lp3 protein 19%2C head-to-tail joining
chromosome	CDS	50579	50935	+	2	prophage Lp3 protein 20
chromosome	CDS	51513	52184	+	0	hypothetical protein

chromosome	CDS	131004	133742	-	0	Cation transport ATPase
chromosome	CDS	134027	134785	+	2	Ontology_term=KEGG_ENZYME:2.4.2.3
chromosome	CDS	134828	135022	-	2	Putative stress-responsive transcriptional regulator
chromosome	CDS	135088	135690	-	1	oxidoreductase (putative)(EC:1.-)
chromosome	CDS	135803	136276	+	2	transcription regulator (putative)
chromosome	CDS	136422	136844	+	0	small heat shock protein
chromosome	CDS	136932	137387	-	0	FIG00750448: hypothetical protein
chromosome	CDS	137384	137515	-	2	Unknown
chromosome	CDS	137967	139199	+	0	Multidrug-efflux transporter%2C major facilitator superfamily (MFS) (TC 2.A.1)
chromosome	CDS	139263	139589	+	0	Transcriptional regulator%2C ArsR family
chromosome	CDS	139654	140247	+	1	Cyanate permease
chromosome	CDS	140234	140848	+	2	Cyanate permease
chromosome	CDS	141092	141940	+	2	oxidoreductase of aldo/keto reductase family%2C subgroup 1
chromosome	CDS	142017	142889	+	0	Aryl-alcohol dehydrogenase related enzyme
chromosome	CDS	143010	143465	+	0	FIG00751510: hypothetical protein
chromosome	CDS	143785	145479	+	1	Predicted oxidoreductase%3B Myosin-crossreactive antigen ortholog
chromosome	CDS	145651	146367	-	1	histone H1
chromosome	CDS	146633	148150	+	2	FIG00743680: hypothetical protein
chromosome	CDS	148219	149328	-	1	Putative NADH-dependent flavin oxidoreductase
chromosome	CDS	149579	150418	-	2	Transmembrane component MtsC of energizing module of methionine-regulated ECF transporter
chromosome	CDS	150420	152120	-	0	Duplicated ATPase component MtsB of energizing module of methionine-regulated ECF transporter

chromosome	CDS	152125	152685	-	1	Substrate-specific component MtsA of methionine-regulated ECF transporter
chromosome	CDS	152764	152931	-	1	FIG00751736: hypothetical protein
chromosome	CDS	153293	154009	+	2	Predicted transcriptional regulators
chromosome	CDS	154154	154483	+	2	FIG00751412: hypothetical protein
chromosome	CDS	154512	154649	+	0	FIG00744409: hypothetical protein
chromosome	CDS	154665	155297	+	0	FIG00744409: hypothetical protein
chromosome	CDS	155322	156134	+	0	FIG00750797: hypothetical protein
chromosome	CDS	156465	156956	+	0	FIG00752778: hypothetical protein
chromosome	CDS	157125	157919	+	0	oxidoreductase%2C short chain dehydrogenase/reductase family
chromosome	CDS	158182	159099	+	1	ABC transporter%2C ATP-binding protein
chromosome	CDS	159099	160721	+	0	FIG00752240: hypothetical protein
chromosome	CDS	160740	161402	+	0	Transcriptional regulator%2C TetR family
chromosome	CDS	161569	163110	+	1	FIG00748417: hypothetical protein
chromosome	CDS	163283	163795	-	2	Substrate-specific component FolT of folate ECF transporter
chromosome	CDS	271202	276226	+	2	Ontology_term=KEGG_ENZYME:2.7.8.12
chromosome	CDS	276625	279903	+	1	Ontology_term=KEGG_ENZYME:2.7.8.12
chromosome	CDS	280241	280804	+	2	UbiX family decarboxylase%2C lactobacillus type
chromosome	CDS	280807	281217	+	1	FIG00750458: hypothetical protein
chromosome	CDS	528752	529897	-	2	Integrase
chromosome	CDS	529946	530854	-	2	hypothetical protein
chromosome	CDS	531201	531884	+	0	prophage Lp4 protein 3%2C phage-like repressor
chromosome	CDS	531938	532204	+	2	prophage Lp4 protein 4

chromosome	CDS	532261	532374	+	1	hypothetical protein
chromosome	CDS	532516	532938	+	1	prophage Lp4 protein 5
chromosome	CDS	532931	533152	+	2	prophage Lp4 protein 6
chromosome	CDS	533202	533945	+	0	prophage Lp4 protein 7%2C DNA replication
chromosome	CDS	533957	535375	+	2	prophage Lp3 protein 8%2C helicase
chromosome	CDS	535986	536360	+	0	prophage Lp4 protein 11%2C DNA replication
chromosome	CDS	536362	536775	+	1	prophage Lp4 protein 12
chromosome	CDS	537821	538147	+	2	Unknown
chromosome	CDS	538491	538745	-	0	hypothetical protein
chromosome	CDS	738816	740342	-	0	Ontology_term=KEGG_ENZYME:2.7.1.30
chromosome	CDS	740647	741321	+	1	FIG00752445: hypothetical protein
chromosome	CDS	741816	742238	+	0	putative arsenate reductase
chromosome	CDS	742506	742736	+	0	Unknown
chromosome	CDS	742904	743416	-	2	putative membrane protein
chromosome	CDS	743532	744977	-	0	drug resistance transporter%2C EmrB/QacA family
chromosome	CDS	745147	746100	-	1	Ontology_term=KEGG_ENZYME:3.6.1.11
chromosome	CDS	746103	748259	-	0	Ontology_term=KEGG_ENZYME:2.7.4.1
chromosome	CDS	748260	749789	-	0	Ontology_term=KEGG_ENZYME:3.6.1.11
chromosome	CDS	750114	750923	-	0	Ontology_term=KEGG_ENZYME:2.7.8.-
chromosome	CDS	751080	751565	+	0	Transcriptional regulator%2C MarR family
chromosome	CDS	1125724	1126266	-	1	Ontology_term=KEGG_ENZYME:2.4.2.7
chromosome	CDS	1126269	1127342	-	0	Guanine-hypoxanthine permease

chromosome	CDS	1127537	1128058	+	2	FIG00746103: hypothetical protein
chromosome	CDS	1128250	1128768	-	1	Histone acetyltransferase HPA2 and related acetyltransferases
chromosome	CDS	1128883	1129719	-	1	Diadenosine tetraphosphatase and related serine/threonine protein phosphatases
chromosome	CDS	1130074	1130496	+	1	Transcriptional regulator%2C AraC family
chromosome	CDS	1130557	1130946	+	1	Transcriptional regulator%2C AraC family
chromosome	CDS	1131064	1131951	+	1	Esterase/lipase
chromosome	CDS	1131970	1133184	+	1	FIG00748805: hypothetical protein
chromosome	CDS	1133177	1134070	+	2	Transcriptional regulator/sugar kinase
chromosome	CDS	1134223	1135548	-	1	Manganese transport protein MntH
chromosome	CDS	1135520	1136494	-	2	Ontology_term=KEGG_ENZYME:4.99.1.1
chromosome	CDS	1136826	1138208	-	0	S-methylmethionine permease
chromosome	CDS	1138322	1139251	+	2	Ontology_term=KEGG_ENZYME:2.1.1.10
chromosome	CDS	1139649	1141148	+	0	Ontology_term=KEGG_ENZYME:2.4.1.52
chromosome	CDS	1141224	1141943	-	0	FIG00747924: hypothetical protein
chromosome	CDS	1142157	1143344	+	0	Ontology_term=KEGG_ENZYME:2.5.1.6
chromosome	CDS	1143481	1144983	+	1	Permeases of the major facilitator superfamily
chromosome	CDS	1145423	1148332	+	2	FIG00753378: hypothetical protein
chromosome	CDS	1506004	1506252	+	1	integral membrane protein
chromosome	CDS	1506324	1506443	+	0	Unknown
chromosome	CDS	1506459	1506725	+	0	FIG028593: membrane protein
chromosome	CDS	1506815	1506928	-	2	hypothetical protein
chromosome	CDS	1506937	1507158	-	1	hypothetical protein

chromosome	CDS	1507236	1507502	-	0	integral membrane protein
chromosome	CDS	1507686	1509125	-	0	Dipeptidase
chromosome	CDS	1509395	1509589	+	2	Ontology_term=KEGG_ENZYME:5.3.2.-
chromosome	CDS	1509662	1510987	+	2	Ontology_term=KEGG_ENZYME:4.1.1.20
chromosome	CDS	1511393	1512310	+	2	Ontology_term=KEGG_ENZYME:2.5.1.74
chromosome	CDS	1512409	1513185	+	1	Potassium voltage-gated channel subfamily KQT%3B possible potassium channel%2C VIC family
chromosome	CDS	1513278	1513577	+	0	FIG00753398: hypothetical protein
chromosome	CDS	1513717	1513989	+	1	lipoprotein precursor (putative)
chromosome	CDS	1514053	1514199	-	1	hypothetical protein
chromosome	CDS	1514423	1515772	+	2	Aminotransferase
chromosome	CDS	1515789	1517288	+	0	Amino acid transporter
chromosome	CDS	1517470	1518279	-	1	Hydrolase (HAD superfamily)
chromosome	CDS	1518427	1519794	+	1	tRNA and rRNA cytosine-C5-methylases
chromosome	CDS	1519963	1520835	+	1	FIG00751131: hypothetical protein
chromosome	CDS	1521329	1522675	+	2	Sugar transporter
chromosome	CDS	1524993	1525973	+	0	Ontology_term=KEGG_ENZYME:5.1.3.3
chromosome	CDS	1526130	1527176	-	0	Ontology_term=KEGG_ENZYME:5.3.3.2
chromosome	CDS	1527248	1528363	-	2	Ontology_term=KEGG_ENZYME:2.7.4.2
chromosome	CDS	1528390	1529367	-	1	Ontology_term=KEGG_ENZYME:4.1.1.33
chromosome	CDS	1529393	1530331	-	2	Ontology_term=KEGG_ENZYME:2.7.1.36
chromosome	CDS	1531176	1533980	+	0	DinG family ATP-dependent helicase YoaA

chromosome	CDS	1534085	1534546	+	2	FIG00742179: hypothetical protein
chromosome	CDS	1534569	1535768	+	0	Ontology_term=KEGG_ENZYME:2.6.1.1
chromosome	CDS	1535786	1537084	+	2	Ontology_term=KEGG_ENZYME:6.1.1.22
chromosome	CDS	1537162	1537887	+	1	Chromosome replication initiation protein DnaD
chromosome	CDS	1538523	1539590	+	0	Methionine ABC transporter ATP-binding protein
chromosome	CDS	1539590	1540261	+	2	Methionine ABC transporter permease protein
chromosome	CDS	1540280	1541137	+	2	Methionine ABC transporter substrate-binding protein
chromosome	CDS	1541228	1541728	+	2	Universal stress protein family
chromosome	CDS	1541700	1541813	+	0	hypothetical protein
chromosome	CDS	1541847	1543532	-	0	Ferric iron ABC transporter%2C permease protein
chromosome	CDS	1543529	1544539	-	2	Ferric iron ABC transporter%2C iron-binding protein
chromosome	CDS	1544552	1545505	-	2	Molybdenum transport ATP-binding protein ModC (TC 3.A.1.8.1)
chromosome	CDS	1545690	1547993	-	0	Ontology_term=KEGG_ENZYME:2.4.1.129,KEGG_ENZYME:3.4.-.-
chromosome	CDS	1548005	1548628	-	2	RecU Holliday junction resolvase
chromosome	CDS	1548743	1549312	+	2	FIG005686: hypothetical protein
chromosome	CDS	1549387	1549728	+	1	Cell division protein GpsB%2C coordinates the switch between cylindrical and septal cell wall synthesis by re-localization of PBP1
chromosome	CDS	1549789	1549911	+	1	hypothetical protein
chromosome	CDS	1550288	1551433	+	2	FIG001721: Predicted N6-adenine-specific DNA methylase
chromosome	CDS	1551471	1551767	+	0	Transcriptional regulator%2C ArsR family
chromosome	CDS	1551808	1551927	-	1	hypothetical protein
chromosome	CDS	1552030	1553394	+	1	FIG00749327: hypothetical protein

chromosome	CDS	1553523	1554275	-	0	Putative hydrolase of the alpha/beta superfamily
chromosome	CDS	1554471	1554878	+	0	FIG00753608: hypothetical protein
chromosome	CDS	1554939	1556279	+	0	FIG00744610: hypothetical protein
chromosome	CDS	1556433	1557512	+	0	Ontology_term=KEGG_ENZYME:3.4.17.13
chromosome	CDS	1557610	1558098	+	1	Protein export cytoplasm protein SecA ATPase RNA helicase (TC 3.A.5.1.1)
chromosome	CDS	1558246	1559592	+	1	FIG00744905: hypothetical protein
chromosome	CDS	1559764	1560732	+	1	Ontology_term=KEGG_ENZYME:3.4.17.13
chromosome	CDS	1560874	1561632	+	1	FIG00750966: hypothetical protein
chromosome	CDS	1561632	1562309	+	0	possible ABC transporter%2C ATP binding component
chromosome	CDS	1562327	1562467	+	2	hypothetical protein
chromosome	CDS	1562464	1563543	-	1	integral membrane protein
chromosome	CDS	1563800	1564594	+	2	hydrolase (putative)
chromosome	CDS	1564752	1565858	+	0	Glycine/D-amino acid oxidases family
chromosome	CDS	1565855	1566241	-	2	Ribonuclease HI%2C Bacillus nonfunctional homolog
chromosome	CDS	1566276	1566662	+	0	FIG00742356: hypothetical protein
chromosome	CDS	1566756	1568411	+	0	Ontology_term=KEGG_ENZYME:6.3.4.3
chromosome	CDS	1568744	1569193	+	2	Ontology_term=KEGG_ENZYME:3.4.23.36
chromosome	CDS	1569213	1570136	+	0	Ontology_term=KEGG_ENZYME:4.2.1.70
chromosome	CDS	1570295	1570819	+	2	Ontology_term=KEGG_ENZYME:2.4.2.9
chromosome	CDS	1570854	1571936	+	0	Ontology_term=KEGG_ENZYME:6.3.5.5
chromosome	CDS	1571933	1574494	+	2	Ontology_term=KEGG_ENZYME:6.3.5.5
chromosome	CDS	1574691	1575272	+	0	Phosphoglycerate mutase family

chromosome	CDS	1575383	1575778	-	2	FIG00748508: hypothetical protein
chromosome	CDS	1575945	1576064	-	0	hypothetical protein
chromosome	CDS	2033709	2035745	-	0	DNA mismatch repair protein MutL
chromosome	CDS	2035913	2038603	-	2	DNA mismatch repair protein MutS
chromosome	CDS	2038631	2039437	-	2	FIG006542: Phosphoesterase
chromosome	CDS	2039562	2041121	-	0	FIG002344: Hydrolase (HAD superfamily)
chromosome	CDS	2041352	2042494	-	2	RecA protein
chromosome	CDS	2042586	2043848	-	0	Competence/damage-inducible protein CinA
chromosome	CDS	2044120	2044704	-	1	Ontology_term=KEGG_ENZYME:2.7.8.5
chromosome	CDS	2044727	2045599	-	2	Transcriptional regulator in cluster with unspecified monosaccharide ABC transport system
chromosome	CDS	2045703	2047004	-	0	FIG009210: peptidase%2C M16 family
chromosome	CDS	2047001	2048266	-	2	FIG001621: Zinc protease
chromosome	CDS	2048354	2048479	+	2	hypothetical protein
chromosome	CDS	2048476	2049828	+	1	Ontology_term=KEGG_ENZYME:2.7.2.4
chromosome	CDS	2136635	2136997	-	2	Lj965 prophage protein
chromosome	CDS	2137011	2142842	-	0	Phage tail length tape-measure protein
chromosome	CDS	2142858	2143061	-	0	Lj965 prophage protein
chromosome	CDS	2143229	2143627	-	2	Lj965 prophage protein
chromosome	CDS	2143727	2144197	-	2	Phage tail fibers
chromosome	CDS	2144212	2144577	-	1	Lj965 prophage protein
chromosome	CDS	2144577	2145128	-	0	Lj965 prophage protein

chromosome	CDS	2145130	2145477	-	1	Lj965 prophage protein
chromosome	CDS	2145477	2145809	-	0	hypothetical protein
chromosome	CDS	2145821	2145997	-	2	hypothetical protein
chromosome	CDS	2146010	2147032	-	2	FIG00743554: hypothetical protein
chromosome	CDS	2147052	2147402	-	0	FIG00748876: hypothetical protein
chromosome	CDS	2147417	2148094	-	2	hypothetical protein
chromosome	CDS	2148267	2148473	-	0	hypothetical protein
chromosome	CDS	2148525	2148725	-	0	hypothetical protein
chromosome	CDS	2148778	2150463	-	1	Phage capsid and scaffold
chromosome	CDS	2150480	2150602	-	2	hypothetical protein
chromosome	CDS	2150610	2150867	-	0	hypothetical protein
chromosome	CDS	2150836	2152344	-	1	Phage portal protein
chromosome	CDS	2152356	2153594	-	0	Phage terminase%2C large subunit
chromosome	CDS	2175574	2177115	-	1	prophage Lp3 protein 18
chromosome	CDS	2177112	2178212	-	0	Phage portal protein
chromosome	CDS	2178213	2178347	-	0	hypothetical protein
chromosome	CDS	2178367	2180070	-	1	prophage Lp3 protein 15%2C terminase large subunit
chromosome	CDS	2180067	2180540	-	0	Phage terminase small subunit
chromosome	CDS	2181236	2181508	-	2	Phage protein
chromosome	CDS	2181618	2181956	-	0	Phage protein
chromosome	CDS	2181943	2182134	-	1	prophage Lp3 protein 10
chromosome	CDS	2182149	2182628	-	0	prophage Lp3 protein 9

chromosome	CDS	2182774	2184168	-		1	prophage Lp3 protein 8%2C helicase
chromosome	CDS	2184168	2184968	-		0	prophage Lp4 protein 7%2C DNA replication
chromosome	CDS	2184982	2185209	-		1	hypothetical protein
chromosome	CDS	2185328	2185471	-		2	prophage Lp3 protein 5
chromosome	CDS	2185479	2185658	-		0	prophage Lp3 protein 4
chromosome	CDS	2185805	2186518	+		2	hypothetical protein
chromosome	CDS	2186589	2187344	+		0	prophage Lp3 protein 1%2C integrase
chromosome	CDS	2187367	2187747	+		1	Phage integrase: site-specific recombinase
chromosome	CDS	2501294	2502481	-		2	FIG00754536: hypothetical protein
chromosome	CDS	2502600	2503262	-		0	HAD superfamily hydrolase
chromosome	CDS	2503476	2504960	+		0	FIG00752719: hypothetical protein
chromosome	CDS	2504957	2505457	+		2	FIG00749636: hypothetical protein
chromosome	CDS	2505516	2506073	-		0	Ribosomal-protein-L7p-serine acetyltransferase
chromosome	CDS	2506163	2507173	-		2	Putative membrane protein YeiH
chromosome	CDS	2507290	2508117	+		1	LysR family transcriptional regulator YeiE
chromosome	CDS	2508202	2509710	-		1	Ontology_term=KEGG_ENZYME:2.4.1.52
chromosome	CDS	2509707	2511251	-		0	Ontology_term=KEGG_ENZYME:2.4.1.52
chromosome	tRNA	741520	741591	+	.		tRNA-Asn-GTT
chromosome	tRNA	741520	741591	+	.		tRNA-Asn-GTT
chromosome	tRNA	1125036	1125107	+	.		tRNA-Gln-TTG
chromosome	tRNA	1125036	1125107	+	.		tRNA-Gln-TTG
chromosome	tRNA	1125130	1125200	+	.		tRNA-Cys-GCA

chromosome	tRNA	1125130	1125200	+	.	tRNA-Cys-GCA
chromosome	tRNA	1125256	1125340	+	.	tRNA-Leu-CAA
chromosome	tRNA	1125256	1125340	+	.	tRNA-Leu-CAA

Additional file 3 – Table S2-3. Genomic islands of *L. plantarum* GB-LP3

Island start	Island end	Length	Method	Gene start	Gene end	Strand	Product
310669	318386	7717	Predicted by at least one method	310669	311328	1	ABC transporter, ATP-binding protein
310669	318386	7717	Predicted by at least one method	311457	312218	-1	immunity protein PlnP, membrane-bound protease CAAX family
310669	318386	7717	Predicted by at least one method	312314	313033	-1	membrane-bound protease, CAAX family
310669	318386	7717	Predicted by at least one method	313155	313892	-1	Aggregation promoting factor
310669	318386	7717	Predicted by at least one method	314947	315582	-1	Aggregation promoting factor
310669	318386	7717	Predicted by at least one method	316207	316740	1	Mobile element protein
310669	318386	7717	Predicted by at least one method	316846	317088	1	FIG00747444: hypothetical protein
310669	318386	7717	Predicted by at least one method	317118	318002	1	FIG00746403: hypothetical protein
310669	318386	7717	Predicted by at least one method	318012	318386	1	Glycine cleavage system H protein
399850	406669	6819	Predicted by at least one method	400378	401553	1	ISEf1, transposase
399850	406669	6819	Predicted by at least one method	401826	403166	1	Three-component quorum-sensing regulatory system, sensor histidine kinase
399850	406669	6819	Predicted by at least one method	403167	403910	1	Three-component quorum-sensing regulatory system, response regulator
399850	406669	6819	Predicted by at least one method	404211	404984	-1	Bacteriocin immunity protein (putative), membrane-bound protease CAAX family

399850	406669	6819	Predicted by at least one method	405083	405241	-1	bacteriocin precursor peptide PlnF (putative)
399850	406669	6819	Predicted by at least one method	405318	405431	-1	bacteriocin precursor peptide PlnE (putative)
399850	406669	6819	Predicted by at least one method	405698	406669	1	Bacteriocin ABC-transporter, ATP-binding and permease component
809958	825734	15776	Predicted by at least one method	809958	811889	1	Stage V sporulation protein K
809958	825734	15776	Predicted by at least one method	812267	812431	1	hypothetical protein
809958	825734	15776	Predicted by at least one method	812673	812876	1	Unknown
809958	825734	15776	Predicted by at least one method	812908	813207	1	transposase, fragment (putative)
809958	825734	15776	Predicted by at least one method	813215	813472	1	Unknown
809958	825734	15776	Predicted by at least one method	813469	813714	1	transposase, fragment (putative)
809958	825734	15776	Predicted by at least one method	813905	814084	1	transposase, fragment
809958	825734	15776	Predicted by at least one method	814210	815613	-1	hypothetical protein
809958	825734	15776	Predicted by at least one method	816029	816145	1	hypothetical protein
809958	825734	15776	Predicted by at least one method	816171	816539	-1	Prophage Lp1 protein 6
809958	825734	15776	Predicted by at least one method	816773	817036	-1	Predicted transcriptional regulators
809958	825734	15776	Predicted by at least one method	817707	818390	1	membrane-bound protease, CAAX family
809958	825734	15776	Predicted by at least one	818610	818954	1	Predicted transcriptional regulator

			method				
809958	825734	15776	Predicted by at least one method	819110	819817	1	ISSth1, transposase (orf1), IS3 family
809958	825734	15776	Predicted by at least one method	819865	820701	1	Mobile element protein
809958	825734	15776	Predicted by at least one method	820778	821902	-1	transport protein
809958	825734	15776	Predicted by at least one method	822365	824866	1	FIG00744318: hypothetical protein
809958	825734	15776	Predicted by at least one method	825372	825734	1	MORN motif family protein
809958	825734	15776	Predicted by at least one method	825724	827571	1	Acyltransferase family
841892	850123	8231	Predicted by at least one method	841892	842263	1	Mobile element protein
841892	850123	8231	Predicted by at least one method	842236	842667	1	Mobile element protein
841892	850123	8231	Predicted by at least one method	842887	843804	1	Integrase
841892	850123	8231	Predicted by at least one method	843815	844324	-1	Type I restriction-modification system, specificity subunit S (EC 3.1.21.3)
841892	850123	8231	Predicted by at least one method	844305	844937	1	Type I restriction-modification system, specificity subunit S (EC 3.1.21.3)
841892	850123	8231	Predicted by at least one method	845314	845844	-1	FIG00750565: hypothetical protein
841892	850123	8231	Predicted by at least one method	845959	849528	-1	FIG00751546: hypothetical protein
841892	850123	8231	Predicted by at least one method	849734	849856	1	hypothetical protein
841892	850123	8231	Predicted by at least one method	849944	850123	1	FIG00743700: hypothetical protein

1096650	1104004	7354	Predicted by at least one method	1096505	1098148	1	Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein OppA (TC 3.A.1.5.1)
1096650	1104004	7354	Predicted by at least one method	1098306	1099235	1	Oligopeptide transport system permease protein OppB (TC 3.A.1.5.1)
1096650	1104004	7354	Predicted by at least one method	1099239	1100273	1	Oligopeptide transport system permease protein OppC (TC 3.A.1.5.1)
1096650	1104004	7354	Predicted by at least one method	1100289	1101368	1	Oligopeptide transport ATP-binding protein OppD (TC 3.A.1.5.1)
1096650	1104004	7354	Predicted by at least one method	1101375	1102340	1	Oligopeptide transport ATP-binding protein OppF (TC 3.A.1.5.1)
1096650	1104004	7354	Predicted by at least one method	1102870	1103259	1	hypothetical protein
1096650	1104004	7354	Predicted by at least one method	1103453	1104004	1	integrase/recombinase, fragment (putative)
1199755	1212903	13148	Predicted by at least one method	1200023	1200454	-1	prophage Lp2 protein 8
1199755	1212903	13148	Predicted by at least one method	1200464	1200844	-1	Phage transcriptional regulator, Cro/CI family
1199755	1212903	13148	Predicted by at least one method	1201115	1201297	1	Phage repressor
1199755	1212903	13148	Predicted by at least one method	1201310	1202119	1	Tec protein
1199755	1212903	13148	Predicted by at least one method	1202116	1202319	1	hypothetical protein
1199755	1212903	13148	Predicted by at least one method	1202319	1202657	1	hypothetical protein
1199755	1212903	13148	Predicted by at least one method	1202771	1203019	1	hypothetical protein
1199755	1212903	13148	Predicted by at least one method	1203022	1203222	1	hypothetical protein

1199755	1212903	13148	Predicted by at least one method	1203401	1203547	1	hypothetical protein
1199755	1212903	13148	Predicted by at least one method	1203547	1204407	1	Phage protein
1199755	1212903	13148	Predicted by at least one method	1204407	1205033	1	Phage-associated recombinase
1199755	1212903	13148	Predicted by at least one method	1205030	1205491	1	Single-stranded DNA-binding protein
1199755	1212903	13148	Predicted by at least one method	1205506	1206198	1	Phage protein
1199755	1212903	13148	Predicted by at least one method	1206240	1206956	1	Phage replication initiation protein
1199755	1212903	13148	Predicted by at least one method	1206956	1207741	1	DNA replication protein
1199755	1212903	13148	Predicted by at least one method	1207877	1208179	1	hypothetical protein
1199755	1212903	13148	Predicted by at least one method	1208182	1208493	1	hypothetical protein
1199755	1212903	13148	Predicted by at least one method	1208497	1208655	1	prophage Lp2 protein 25
1199755	1212903	13148	Predicted by at least one method	1208658	1209137	1	prophage LambdaSo, DNA modification methyltransferase, putative
1199755	1212903	13148	Predicted by at least one method	1209403	1209585	1	hypothetical protein
1199755	1212903	13148	Predicted by at least one method	1209597	1210028	1	hypothetical protein
1199755	1212903	13148	Predicted by at least one method	1210382	1211164	1	hypothetical protein
1199755	1212903	13148	Predicted by at least one method	1211431	1211613	1	hypothetical protein
1199755	1212903	13148	Predicted by at least one	1212661	1212903	1	hypothetical protein

			method				
1219255	1235125	15870	Predicted by at least one method	1218851	1219258	1	Phage protein
1219255	1235125	15870	Predicted by at least one method	1219255	1219677	1	FIG00775153: hypothetical protein
1219255	1235125	15870	Predicted by at least one method	1219692	1220303	1	Phage major tail protein
1219255	1235125	15870	Predicted by at least one method	1220395	1220709	1	hypothetical protein
1219255	1235125	15870	Predicted by at least one method	1220733	1220954	1	hypothetical protein
1219255	1235125	15870	Predicted by at least one method	1220973	1225433	1	Phage tail length tape-measure protein
1219255	1235125	15870	Predicted by at least one method	1225437	1226258	1	Phage protein
1219255	1235125	15870	Predicted by at least one method	1226278	1231299	1	Phage endopeptidase
1219255	1235125	15870	Predicted by at least one method	1231317	1231808	1	hypothetical protein
1219255	1235125	15870	Predicted by at least one method	1231810	1232244	1	hypothetical protein
1219255	1235125	15870	Predicted by at least one method	1232274	1232651	1	hypothetical protein
1219255	1235125	15870	Predicted by at least one method	1232651	1232923	1	hypothetical protein
1219255	1235125	15870	Predicted by at least one method	1232923	1233207	1	hypothetical protein
1219255	1235125	15870	Predicted by at least one method	1233207	1234112	1	Lysin
1219255	1235125	15870	Predicted by at least one method	1234687	1235112	-1	lipoprotein precursor (putative)

1440603	1459102	18499	Predicted by at least one method	1440603	1440848	1	Acyl carrier protein
1440603	1459102	18499	Predicted by at least one method	1441265	1441960	1	Ribonuclease III (EC 3.1.26.3)
1440603	1459102	18499	Predicted by at least one method	1441991	1445548	1	Chromosome partition protein smc
1440603	1459102	18499	Predicted by at least one method	1445567	1447105	1	Signal recognition particle receptor protein FtsY (=alpha subunit) (TC 3.A.5.1.1)
1440603	1459102	18499	Predicted by at least one method	1447283	1447630	1	Signal recognition particle associated protein
1440603	1459102	18499	Predicted by at least one method	1447653	1449107	1	Signal recognition particle, subunit Ffh SRP54 (TC 3.A.5.1.1)
1440603	1459102	18499	Predicted by at least one method	1449204	1449476	1	SSU ribosomal protein S16p
1440603	1459102	18499	Predicted by at least one method	1449494	1449751	1	KH domain RNA binding protein YlqC
1440603	1459102	18499	Predicted by at least one method	1449912	1450436	1	16S rRNA processing protein RimM
1440603	1459102	18499	Predicted by at least one method	1450436	1451176	1	tRNA (Guanine37-N1) -methyltransferase (EC 2.1.1.31)
1440603	1459102	18499	Predicted by at least one method	1451334	1451690	1	LSU ribosomal protein L19p
1440603	1459102	18499	Predicted by at least one method	1452126	1452446	1	FIG00748957: hypothetical protein
1440603	1459102	18499	Predicted by at least one method	1453074	1454873	1	FIG00749174: hypothetical protein
1440603	1459102	18499	Predicted by at least one method	1454909	1459102	1	FIG00749174: hypothetical protein
2150867	2173355	22488	Predicted by at least one method	2150836	2152344	-1	Phage portal protein
2150867	2173355	22488	Predicted by at least one	2152356	2153594	-1	Phage terminase, large subunit

			method				
2150867	2173355	22488	Predicted by at least one method	2153584	2154462	-1	Phage terminase small subunit
2150867	2173355	22488	Predicted by at least one method	2154502	2154765	-1	Phage protein
2150867	2173355	22488	Predicted by at least one method	2154978	2155601	-1	hypothetical protein
2150867	2173355	22488	Predicted by at least one method	2156516	2156683	-1	hypothetical protein
2150867	2173355	22488	Predicted by at least one method	2156925	2157401	-1	hypothetical protein
2150867	2173355	22488	Predicted by at least one method	2157534	2157701	-1	prophage Lp1 protein 32
2150867	2173355	22488	Predicted by at least one method	2157698	2158138	-1	prophage Lp1 protein 31
2150867	2173355	22488	Predicted by at least one method	2158329	2158751	-1	Phage protein
2150867	2173355	22488	Predicted by at least one method	2159304	2159417	-1	hypothetical protein
2150867	2173355	22488	Predicted by at least one method	2159410	2159790	-1	prophage Lp2 protein 24
2150867	2173355	22488	Predicted by at least one method	2159787	2160305	-1	prophage Lp1 protein 23
2150867	2173355	22488	Predicted by at least one method	2160302	2160589	-1	prophage Lp1 protein 21
2150867	2173355	22488	Predicted by at least one method	2160586	2161539	-1	prophage Lp1 protein 20
2150867	2173355	22488	Predicted by at least one method	2161622	2162596	-1	FIG00602239: hypothetical protein
2150867	2173355	22488	Predicted by at least one method	2162608	2163138	-1	Phage protein

2150867	2173355	22488	Predicted by at least one method	2163156	2163269	-1	hypothetical protein
2150867	2173355	22488	Predicted by at least one method	2163631	2163888	-1	prophage Lp2 protein 13
2150867	2173355	22488	Predicted by at least one method	2164210	2164515	-1	prophage Lp2 protein 12
2150867	2173355	22488	Predicted by at least one method	2165113	2165475	1	prophage Lp1 protein 8
2150867	2173355	22488	Predicted by at least one method	2165923	2166858	1	Serine transporter
2150867	2173355	22488	Predicted by at least one method	2166997	2167122	1	hypothetical protein
2150867	2173355	22488	Predicted by at least one method	2167202	2167570	1	Prophage Lp1 protein 6
2150867	2173355	22488	Predicted by at least one method	2167575	2167688	1	hypothetical protein
2150867	2173355	22488	Predicted by at least one method	2168170	2169249	1	Streptococcal hemagglutinin protein
2150867	2173355	22488	Predicted by at least one method	2169276	2169839	1	FIG00627979: hypothetical protein
2150867	2173355	22488	Predicted by at least one method	2170499	2171593	1	hypothetical protein
2150867	2173355	22488	Predicted by at least one method	2171883	2173004	1	Integrase
2188496	2198126	9630	Predicted by at least one method	2188496	2188768	1	FIG00746198: hypothetical protein
2188496	2198126	9630	Predicted by at least one method	2188817	2189302	-1	Transcriptional regulator
2188496	2198126	9630	Predicted by at least one method	2189429	2189716	1	Mobile element protein
2188496	2198126	9630	Predicted by at least one method	2189740	2190618	1	Mobile element protein

2188496	2198126	9630	Predicted by at least one method	2190679	2192655	-1	LepB
2188496	2198126	9630	Predicted by at least one method	2192755	2195586	-1	FIG00753378: hypothetical protein
2188496	2198126	9630	Predicted by at least one method	2195842	2195973	-1	hypothetical protein
2188496	2198126	9630	Predicted by at least one method	2196003	2196134	-1	hypothetical protein
2188496	2198126	9630	Predicted by at least one method	2196213	2196344	-1	Unknown
2188496	2198126	9630	Predicted by at least one method	2196376	2196519	-1	hypothetical protein
2188496	2198126	9630	Predicted by at least one method	2196552	2196683	-1	hypothetical protein
2188496	2198126	9630	Predicted by at least one method	2196716	2196847	-1	hypothetical protein
2188496	2198126	9630	Predicted by at least one method	2196880	2197011	-1	Unknown
2188496	2198126	9630	Predicted by at least one method	2197042	2197185	-1	hypothetical protein
2188496	2198126	9630	Predicted by at least one method	2197357	2197488	-1	hypothetical protein

This chapter will be published elsewhere
as a partial fulfillment of Soomin Jeon's Ph.D program.

Chapter 3. Comparative genomic analysis of *Lactobacillus delbrueckii* subsp. *bulgaricus* and *Limosilactobacillus fermentum* with elevated GC contents among lactic acid bacteria

3.1. Abstract

GC content is considered a result of the adaptation to various environments in Bacteria. Since thymine occurrence due to cytosine deamination is frequently observed in nature, a bacterium with high GC contents is very unusual. This study aimed to discover a species with high GC contents among the lactic acid bacteria and understand the evolution of lactic acid bacteria. *Lactobacillus bulgaricus* and *Limosilactobacillus fermentum* were identified as having high GC contents compared to the genome size. In a comparative analysis on genetic factors whose correlations with GC contents were indicated in previous studies, no difference was found. In both species, an increase of the proportion of codons, whose third position is guanine or cytosine, was verified and the codon usage is less random than other lactic acid bacteria by comparing codon bias and the information entropy. Categorizing isolation source suggests that the elevated GC content results from an adaptation to a nutrient-rich environment. Our study specified two species with elevated GC content and identified that the elevated GC content was associated with CDS increasing synonymous mutations. We hope that our results will provide an understanding of the GC content and bacterial evolution of lactic acid bacteria.

Keywords

Bacterial evolution, Codon usage, GC contents, *Lactobacillus bulgaricus*, *Limosilactobacillus fermentum*

3.2. Introduction

Since guanine binds to cytosine and adenine binds to thymine in DNA helix according to Chargaff's rules, the amount of guanine-cytosine and adenine-thymine remain constant in a genome. GC content is defined as the proportion of guanine and cytosine in the genome sequence. In the bacterial kingdom, it is reported that GC contents have a wide range from 13 % to 75 % (Almpanis, Swain et al. 2018). The lowest GC content was reported as 13.5 % in *Zinderia insecticola* CARI, while the highest GC content was found in *Anaeromyxobacter dehalogenans* 2CP-C with a GC content of 74.9 %. Many researchers have declared that the diversity of GC contents in bacteria is a result of the adaptation to various environments (Hildebrand, Meyer et al. 2010). Since the G-C bond has one more hydrogen bond than the A-T bond consisting of two hydrogen bonds, it has been suggested that the bacteria with high GC contents will have heat-stable characteristics (Basak, Mukhopadhyay et al. 2010). Conversely, the opinion that high GC content is not an adaptation to high temperature has been claimed, and it remains a controversial issue (Hurst and Merchant 2001). It was also reported that the GC content depends on presence of free oxygen in environment (Bohlin, Snipen et al. 2010). Over decades, a number of researches have been conducted focusing on the idea of the relationship between GC content and genetic factors. As a result, several genetic factors have been pointed out to be closely related to GC content: genome size, the number of RNA and

plasmid, horizontal gene transfer, and prophage (Hayek 2013, Almpanis, Swain et al. 2018). Currently, a correlation between genome size and GC content is highlighted as the primary target to understand (Mitchell and communications 2007).

The family *Lactobacillaceae* is a representative lactic acid bacteria and is a gram-positive, non-spore-forming, and catalase-negative taxon (Salvetti, Torriani et al. 2012, Zheng, Wittouck et al. 2020). This family inhabits many fermented foods and gastrointestinal tracts (GIT) of animals. Functionally, it has an inhibitory effect on pathogenic bacteria by producing antibacterial substances such as bacteriocin and lactic acid. Recent studies have revealed that when the family dominates the human intestine, it has beneficial effects such as lowering cholesterol and alleviating inflammatory bowel disease and cancer (Niedzielin, Kordecki et al. 2001, Choi, Kim et al. 2006, Fuentes, Lajo et al. 2013). The GC contents of lactic acid bacteria have been reported from 32 to 54 % according to their genome size (Zheng, Wittouck et al. 2020). Exceptionally, high GC contents were observed in the genomes of *Lactobacillus delbruekii* subsp. *bulgaricus* (*L. bulgaricus*) in spite of its small genome size. According to the study by M van de Guchte et al., distinguishable GC contents of *L. bulgaricus* were offered by the result of the adaptation in a lactose-rich environment (van de Guchte, Penaud et al. 2006). As many pieces of researches have been conducted over the whole

bacterial kingdom, understanding the GC content of lactic acid bacteria is rather limited.

In this study, we aimed to detect a species with high GC content in lactic acid bacteria and provide an understanding of the relationship between bacterial evolution and GC contents. In order to compare GC contents and genetic factors, genomes belonging to the genera *Lactobacillus*, *Lacticaseibacillus*, *Lactiplantibacillus*, *Lactiplantibacillus*, *Ligiactobacillus*, and *Limosilactobacillus* were comparatively analyzed. We hope that this study will allow closer observation of microbial evolution and GC content.

3.3. Materials and Methods

3.3.1. Data collection and construction of a phylogenetic tree

For genome comparison, 375 genomes belonging to the *Lactobacillaceae* family were downloaded from the NCBI database. For comparative analysis, 6 genera of *Lactobacillus*, *Lacticaseibacillus*, *Lactiplantibacillus*, *Lactiplantibacillus*, *Ligiactobacillus*, and *Limosilactobacillus* were used (Table 1).

Table 3-1. Genetic information for comparative analysis

Genus	Species	Abb.	n	Avg. GC	Std. GC	Avg. size	Std. size
Lacticaseibacillus	*casei*	*Lcb. casei*	4	47.536	0.728	2,983,657	101,281
Lacticaseibacillus	*paracasei*	*Lcb. paracasei*	47	46.346	0.097	3,101,736	100,756
Lacticaseibacillus	*rhamnosus*	*Lcb. rhamnosus*	28	46.726	0.048	3,000,085	71,224
Lactiplantibacillus	*pentosus*	*Lpb. pentosus*	5	46.250	0.110	3,663,991	0
Lactiplantibacillus	*plantarum*	*Lpb. plantarum*	146	44.471	5.293	3,320,604	125,196
Lactobacillus	*acetotolerans*	*L. acetotolerans*	4	36.675	0.154	1,687,557	11,924
Lactobacillus	*acidophilus*	*L. acidophilus*	8	34.704	0.014	1,687,557	35,200
Lactobacillus	*amylovorus*	*L. amylovorus*	4	38.031	0.143	1,687,557	109,303
Lactobacillus	*bulgaricus*	*L. bulgaricus*	11	49.686	0.107	1,687,557	80,078
Lactobacillus	*crispatus*	*L. crispatus*	8	37.074	0.156	1,687,557	152,713
Lactobacillus	*gasseri*	*L. gasseri*	7	35.020	0.112	1,687,557	91,871
Lactobacillus	*helveticus*	*L. helveticus*	21	36.988	0.184	1,687,557	87,903
Lactobacillus	*iners*	*L. iners*	7	33.196	0.115	1,687,557	87,186
Lactobacillus	*johnsonii*	*L. johnsonii*	16	34.717	0.221	1,687,557	94,637
Ligilactobacillus	*salivarius*	*Lglb. salivarius*	10	33.009	0.118	2,102,476	144,036
Limosilactobacillus	*fermentum*	*Lm. fermentum*	29	51.369	0.478	2,140,861	122,684
Limosilactobacillus	*reuteri*	*Lm. reuteri*	20	38.920	0.178	2,134,654	114,346

To avoid a bias arising such as a sequencing error or selection of the unusual object, we used only species with three or more complete genomes for the analysis. To provide an evolutionary relationship of genomes used for comparative analysis, we constructed a phylogenetic tree by 16S rRNA genes. RNAmmer1.2 searched 16S rRNA sequence from the bacterial genomes, and MEGA-X (Version 10.2.2) generated the tree (Lagesen, Hallin et al. 2007, Kumar, Stecher et al. 2018). The maximum-likelihood approach with a bootstrap value of 1000 was used to construct the tree.

3.3.2. Genome comparison of *Lactobacillaceae* family

Comparison between GC content and genome size was performed using a series of Python scripts and the ggplot2 package of R studio (Wickham 2011). Pearson correlation analysis was conducted to understand the association between GC content and genome size. The number of CDSs, chromosomes, plasmids, and structural RNAs were counted from genomic data. The prophage gene was discovered by PhiSpy software (Version 4.2.19) (Akhter, Aziz et al. 2012). GC content of each gene and frequency of guanine and cytosine at each nucleotide position within the codon were calculated from the complete genome sequences by serial Python scripts.

3.3.3. Analysis of codon usage and amino acid pattern

For comparative analysis, three groups were set up as experimental and control groups. The experimental group consisted of B (*L. bulgaricus*) and F (*Lm. fermentum*), and other lactic acid bacteria (other LAB) were assigned as a control group. Among the LAB genomes listed in Table 1, sequences with a genome size between 1.5 Mb and 2.5 Mb were selected as a control. Codon usage was calculated from CDS fasta files. The orthologous gene set was found by OrthoFinder using only the genomes belonging to the same genus (Emms and Kelly 2019), and the detected orthologous gene set was aligned by ClustalW 2.1 (Thompson, Higgins et al. 1994).

Relative entropy of codon usage was presented using Kullback–Leibler divergence (D_{KL}) (Bohlin, van Passel et al. 2012, Van Erven and Harremos 2014). D_{KL} is a measure of the difference between two discrete probability mass functions. The D_{KL} for the sequence s is given as:

$$D_{KL}(s) = \sum_{i=1}^{256} O\,(z_i|s) \log\left(\frac{O(z_i|s)}{E(z_i|s)}\right)$$

s means a DNA sequence in the equation, and z is a possible tetramer. $O(z_i|s)$ is an observed tetranucleotides frequency, $E(z_i|s)$ is the expected frequencies of tetranucleotides from the DNA sequence.

3.3.4. Calculation of relative synonymous codon usage and effective number of codon

Relative synonymous codon usage (RSCU) is the ratio of the observed frequency of codons to the expected frequency calculated with an assumption that all codons encoding the same amino acid appeared equally. It is used for the comparison of observed synonymous codon usage variation. The equation is as follow:

$$RSCU_i = \frac{X_i}{\frac{1}{N_i}\sum_{j=1}^{n_i} X_j}$$

where X_i is the observed number of the ith codon for the jth amino acid, which has n_i kinds of different synonymous codons. If an RSCU value is 1, the codon is not biased. Meanwhile, codons with RSCU values of > 1.6 or < 0.6 are considered to be overrepresented or underrepresented, respectively. The RSCU value was computed using the *cusp* package included in EMBOSS-6.6.0.

The effective number of codon usage (ENC) is affected by the strength of codon bias, regardless of the number of amino acids and gene length. The range of ENC values is from 20 to 61. If the value is close to 61, it means that all synonymous codons appear. In contrast, if the value is close to 20, it indicates an extreme bias towards the use of only one codon for each amino acid

$$\mathrm{ENC} = 2 + \frac{9}{\overline{F_2}} + \frac{1}{\overline{F_3}} + \frac{5}{\overline{F_4}} + \frac{3}{\overline{F_6}}$$

where F_i $(i = 2,3,4,6)$ represents the mean value for F_i and i represents i-fold degenerate amino acids. F_i was computed with the following equation:

$$F_i = \frac{n\sum_i^n \left(\frac{n_i}{n}\right)^2 - 1}{n-1}$$

where n is the number of the total occurrences of the codon for that amino acid, n_i is the total number of the ith codon for the corresponding amino acid. The ENC was calculated by the *seqall* package of EMBOSS-6.6.0. The ENC-GC3 plot was generated to identify factors influencing codon usage variation. The expected ENC value for each GC3 was calculated using the following equation:

$$ENC_{expected} = 2 + s + \frac{29}{s^2 + (1-s)^2}$$

3.4.5. Identification of specific genes found in *L. bulgaricus* and *Lm. fermentum*

We performed gene annotation and comparative analysis to detect a genetic factor related to increased GC contents. Since bacteria have different genetic contents depending on species, the species included in the same genus were compared. For *Lactobacillus* analysis, the genomes of 4 *L. amylovorus*, 7 *L. iners*, 14 *L. johnsonii*, *7 L. crispatus*, 8 *L. bulgaricus*, 2 *L. paragasseri*, 8 *L. acidophilus*, 5 *L. gasseri*, 19 *L. helveticus*, 3 *L. kefiranofaciens*, 1 *L. gallinarum*, and 1 *L. amylolyticus* were used. To

compare *Limosilactobacillus* genus, 24 *Lm. fermentum*, 18 *Lm. reuteri*, 2 *Lm. mucosae*, 1 *Lm. gastricus*, 1 *Lm. oris*, 1 *Lm. pontis*, 1 *Lm. vaginalis*, and 1 *Lm. frumenti* were utilized. Bacterial genomes were annotated by Prokka (Seemann 2014). The gene with the highest and the lowest copies was selected as the different genes. In order to select genes with more than one copy than the control group, the average of the experimental group with a difference of 0.5 or less from the highest/lowest values was excluded. Subsequently, functional protein categorization was performed to understand the function of the discovered species-specific gene group. For annotation, COG classification was used through EGGNOG-mapper (Huerta-Cepas, Forslund et al. 2017).

3.4.6. Statistical analysis

All comparative analysis was performed by analysis of variance (ANOVA) to identify a significant difference among comparative groups. After ANOVA, the Bonferroni method was used for *post-hoc*. The FDR correction method was used for multiple testing issues.

3.4. Results

3.4.1. Species identification with high GC contents

The evolutionary relationship of the *Lactobacillaceae* family was investigated by generating a phylogenetic tree (Figure 1). Strains were well clustered as the same species except for four genomes, *Lpb. plantarum* 5-2, *Lpb. plantarum* 16, *Lglb. salivarius* CECT5713, and *Lcb. helveticus* H10. *Lactiplantibacillus* was clustered together with *Lglb. salivarius* and *Lacticaseibacillus*, and *Limosilactobacillus* showed a close evolutionary relationship with *Lactobacillus*, *Lcb. helveticus* and *Lcb. johnsonii*.

Figure 3-1. Phylogenetic tree using 16S rRNA

The tree based on 16S rRNA sequences for 375 species was constructed using the Maximum likelihood method. Each genus was distinguished by color. For example, species names under *Lactiplantibacillus* are colored black. Similarly, pink-colored names assign *Lacticaseibacillus*, and green-colored names correspond to *Lactobacillus*. Purple and red colors represent *Ligilactobacillus* and *Limosilactobacillus*, respectively.

To find a species with a high GC content, we plotted a scatter plot that shows relationships between the GC content and the genome (Figure 2). Strains belonging to the same species were positioned similarly. The density of genome size near 2.0 Mb and 3.0 Mb showed high, and density of GC content near 45% was dominant. A weak positive correlation was found between genome size and GC contents in *Lactobacillaceae* with a slope of 3.8792. In the correlation graph, *L. bulgaricus* and *Limosilactobacillus fermentum* (*Lm. fermentum*) showed higher GC contents than the trend line.

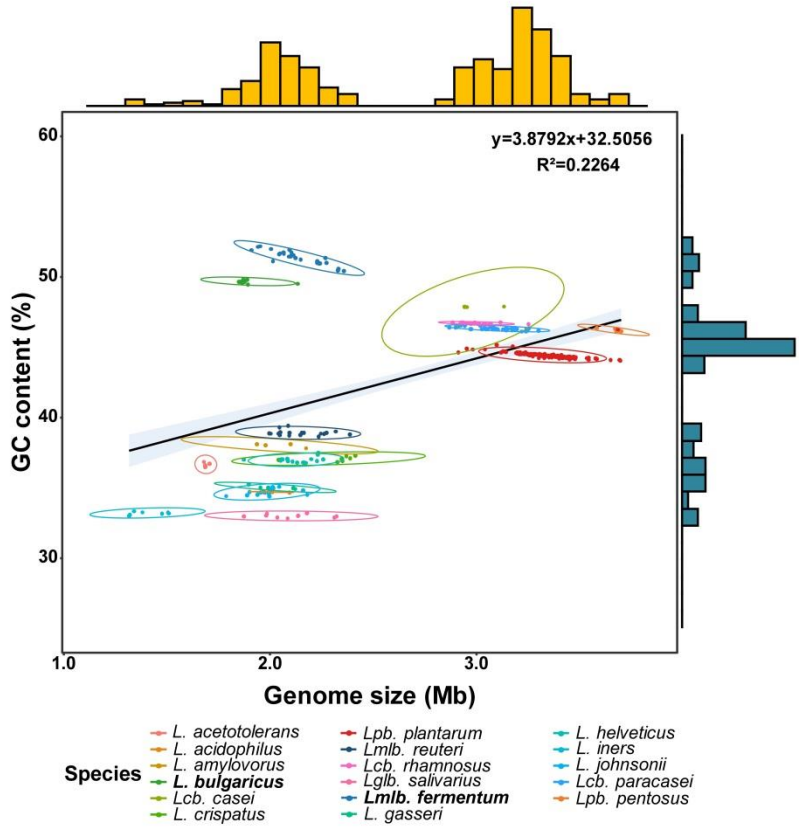


Figure 3-2. Relationship between genome size and GC contents for genomes of lactic acid bacteria

Genome size and GC contents were compared using a scatter plot. Each species was presented in different colors. Density was displayed by a bar chart composed of the yellow-colored x-axis and turquoise-colored y-axis.

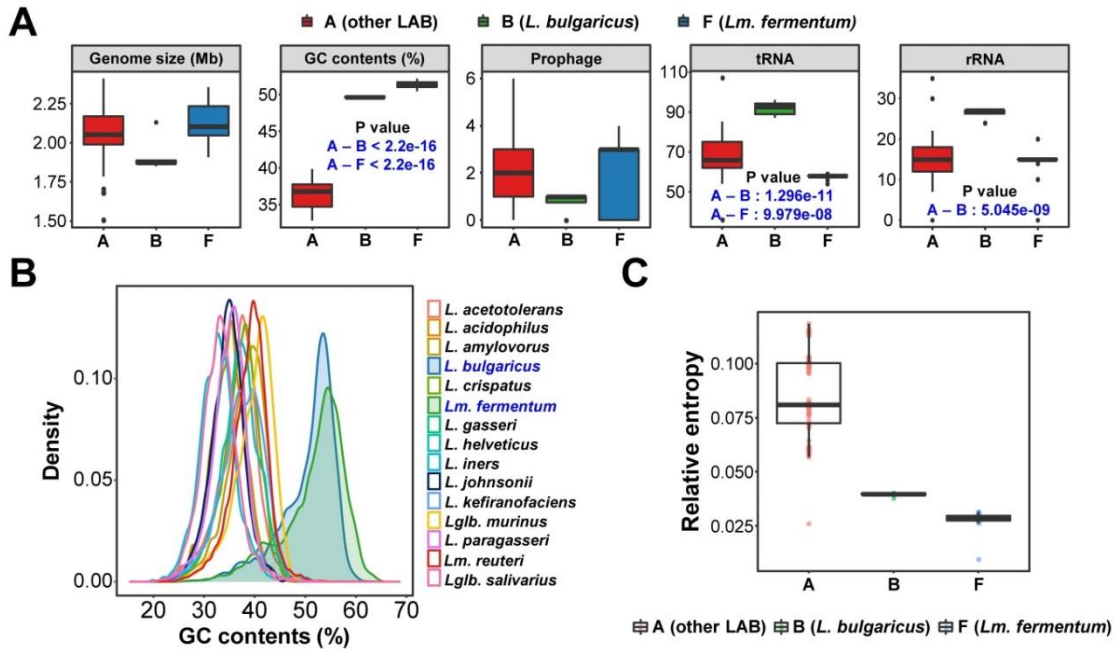


Figure 3-3. Comparison of potential genetic factors related to GC contents

(A) Genetic content comparison of lactic acid bacteria with 1.5-2.5Mb of genome size. A significant difference between comparison groups is denoted by blue-colored bold text with P-value limited under 2.2e-16. **(B)** Gene density by GC contents. Two species with peaks between 50 and 60 are indicated in bold. **(C)** Kullback–Leibler divergence (D_{KL}) of comparison groups. The value was plotted using a box plot, and each value was showed a point.

3.4.2. Comparison of genetic factors

In order to understand the effect of the high GC content in *L. bulgaricus* and *Lm. fermentum*, the genetic factors pointed out in the previous studies were analyzed by groups (Figure 3A and S1A). There was no difference in genome size within each group, but it was confirmed that *L. bulgaricus* and *Lm. fermentum* showed high GC content. At the same time, it was confirmed that CDS of both groups showed higher GC contents than other LAB. Compared with other LAB, *L. bulgaricus* showed more rRNA and tRNA than structural RNAs, but *Lm. fermentum* had only a smaller number of tRNA. On the other hand, there were no differences in the number of chromosomes, plasmids, prophages, ncRNAs, and tmRNAs by groups. Comparing the GC content of the whole genome and CDS, all three groups showed higher GC contents in the CDS than in the whole genome sequence. The increase was higher in *L. bulgaricus* and *Lm. fermentum* as 1.68 and 1.23, while the increase in other LAB was 0.81. To identify the specific genetic factor with higher GC contents, we compared each gene's GC content density distribution (Figure 3B). In *L. bulgaricus* and *Lm. fermentum*, genes were distributed with overturned V-shaped curves at higher GC contents than other species. In the previous study presented on the high GC content of *L. bulgaricus*, it was suggested that the third nucleotide of the codon had a high GC ratio (van de Guchte, Penaud et al. 2006). Based on this result, the codon of CDS for each group was divided

by nucleotide position (GC1, GC2, and GC3, respectively), and the nucleotide composition was compared (Table 2). Significant increases in GC ratio were observed at all positions in *L. bulgaricus* and *Lm. fermentum*. Compared with other LAB, the most significant difference was observed in GC3, followed by GC1 and GC2 in order.

Table 3-2. Statistic values by nucleotide position of codon

Average of GC ratio by nucleotide position			
	GC1	GC2	GC3
Other LAB	47.63 ± 0.19	34.08 ± 0.07	28.96 ± 0.43
L. bulgaricus	53.39 ± 0.05**	36.67 ± 0.16**	64.01 ± 0.21**
Lm. fermentum	55.89 ± 0.09**	38.20 ± 0.13**	63.78 ± 0.28**
Probability of non-synonymous mutation by nucleotide position of codon			
	1st nucleotide	2nd nucleotide	3rd nucleotide
Probability	91.67%	97.62%	32.29%

*: P-value under 0.05, **: P-value under 0.01. All presented comparisons were performed with other LAB groups, and all values were shown as average ± standard error of the average.

3.4.3. Comparison of codon and amino acid patterns

According to the central dogma, the difference in the GC ratio may cause changes in codon usage and amino acid composition. In order to verify the hypothesis, a comparison of codon usage and amino acid composition was conducted. In the comparison of codon usage, a difference in frequency was found in 60 codons except for start codon (ATG), stop codons (TGA, TAG, and TAA), and tryptophan (TGG, it is a single codon for tryptophan.) (Figure S2). In other LAB, codon frequency with A or T at the third position of codon was relatively higher, whereas codon with G or C was more used in *L. bulgaricus* and *Lm. fermentum*. However, the codon usage patterns were inconsistent between *L. bulgaricus* and *Lm. fermentum*. To identify codon preference by encoding an amino acid, we compared the codon ratio in identical amino acids (Figure 4). In *L. bulgaricus* and *Lm. fermentum*, the ratio of codons whose third base is G and C was increased except for stop codons. In the amino acid composition analysis, the frequencies of amino acids were different between the comparative groups except for alanine (Figure S3). Higher frequencies of glutamate and glycine were observed in *L. bulgaricus* and *Lm. fermentum* than other LAB and frequencies of isoleucine and methionine were lower than other LAB. Since different codon frequencies may be affected by the type of encoded CDS, we compared codon usage and amino acid composition using the orthologous gene set, and the G and C ratio at the third nucleotide was

higher in *L. bulgaricus* and *Lm. fermentum* than other LAB (Figure S4). In addition, in order to check whether the difference in GC3 affected the amino acid composition, the sum of the codons that encode the same amino acid but have different third nucleotide was compared (Figure S5). In the statistical analysis, a significant difference was not found.

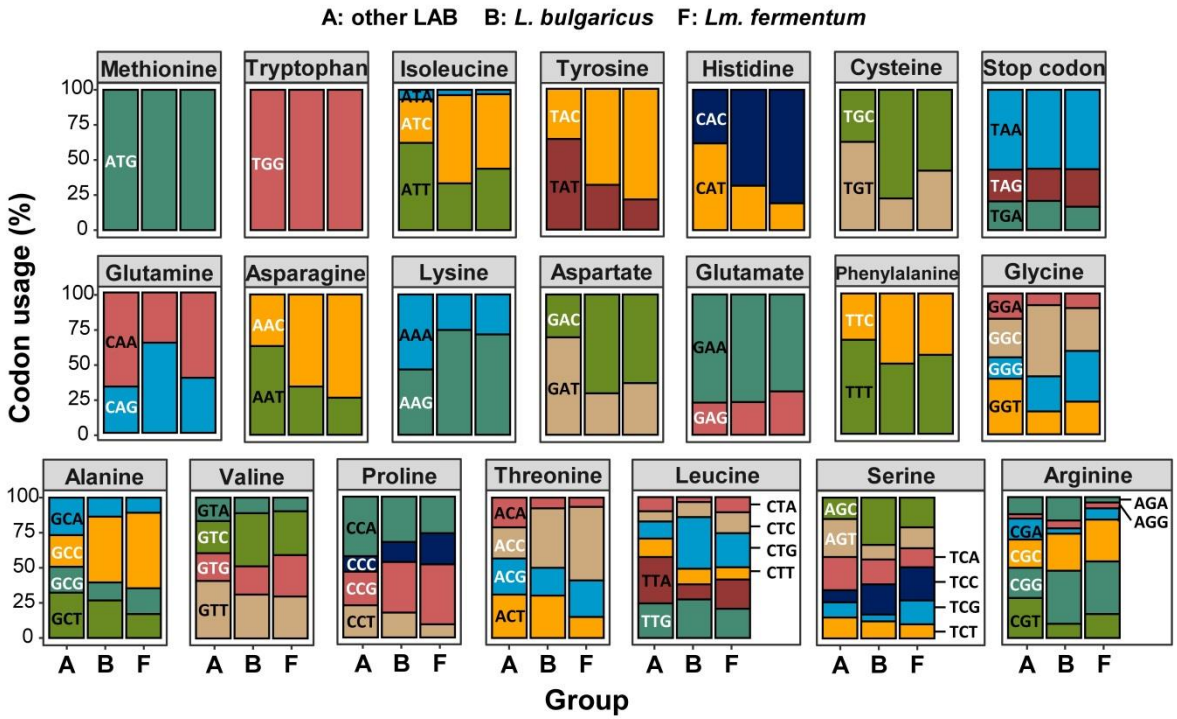


Figure 3-4. Comparison of codon usage by amino acid

The y-axis shows the codon ratio, and the x-axis indicates three comparison groups. Each color bar represents the ratio of each codon.

3.4.4. Analysis of codon usage bias

The codon usage characteristics of lactic acid bacteria were analyzed with statistical factors. First, RSCU was measured to compare observed synonymous codon usage variation (Figure 5). In other LAB group, 9 and 5 codons were overrepresented or underrepresented, respectively. In overrepresented codons, the third nucleotide of the codon was either adenine or thymine, and 4 out of 5 underrepresented codons had guanine or cytosine at the third nucleotide of the codon. In *L. bulgaricus* and *Lm. fermentum* groups, a greater number of codons were overrepresented or underrepresented than in other LAB. Most of the third nucleotide of the overrepresented codon was cytosine (*L. bulgaricus* was 61.54% and *Lm. fermentum* was 66.67%), and one of the underrepresented codons was often adenine (*L. bulgaricus* and *Lm. fermentum* had 60.00% and 66.67%, respectively). Although *Lm. fermentum* and *L. bulgaricus* showed similar codon preferences for arginine, isoleucine, and threonine, there was no difference for other amino acids.

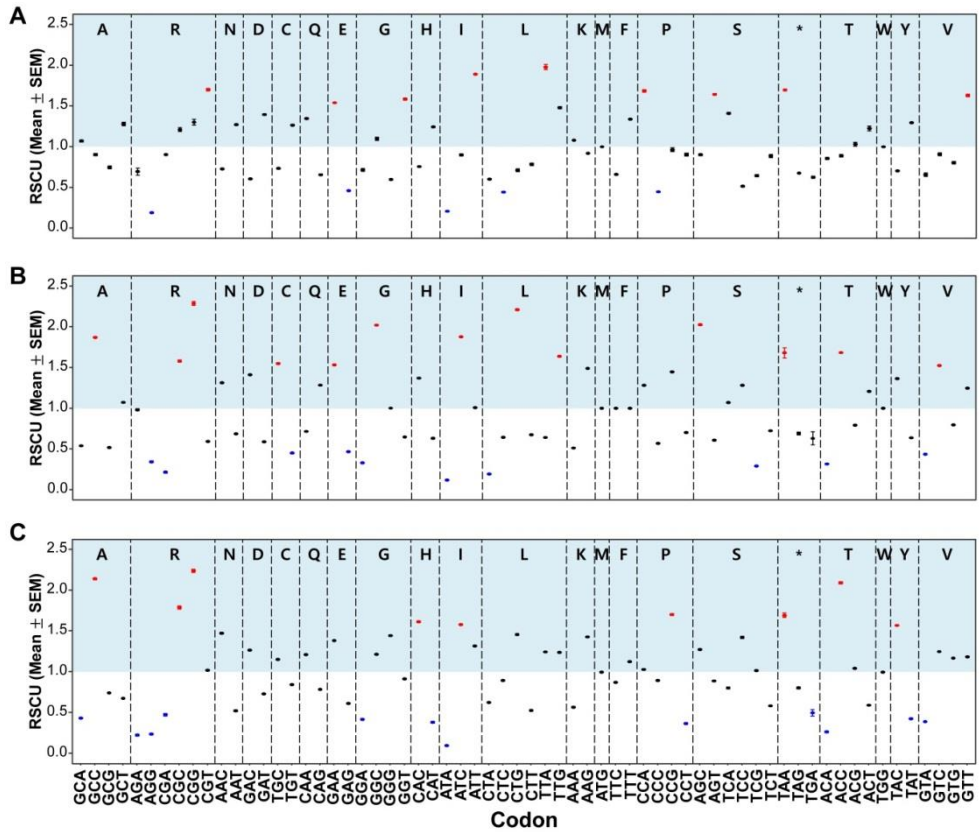


Figure 3-5. RSCU analysis of the diverse codons

The RSCU values are represented on the y-axis with SEM, while the codon families for each amino acid are denoted on the x-axis by groups: **(A)** is other LAB, and **(B)** and **(C)** shows RSCU values of *L. bulgaricus* and *Lm. fermentum*, respectively. Highly preferred codons (RSCU > 1.5) are highlighted in red, and unpreferred codons are in blue with RSCU < 0.5.

ENC was calculated to evaluate the overall codon usage bias (Figure 6A). If the ENC value is close to 60, it means the weak codon bias (Wright 1990). If the calculated ENC value is below the expected ENC-GC3 curve, the selection is interpreted as the main factor of the codon bias. Likewise, if the calculated ENC is positioned under the expected ENC-GC3 curve, it indicates that mutation is the main factor of codon bias. No statistical difference was found in calculated ENC values among the groups in our data. The calculated ENC values of all groups were below the expected ENC-GC3 curve, indicating that selection is the main factor of codon bias. To compare the diversity of codon usage patterns in each group, we compared the information entropy of codon usage by calculating the Kullback-Leibler divergence value (Figure 3C and S1B). It was found that *L. bulgaricus* and *Lm. fermentum* had significantly lower entropy values than other LAB.

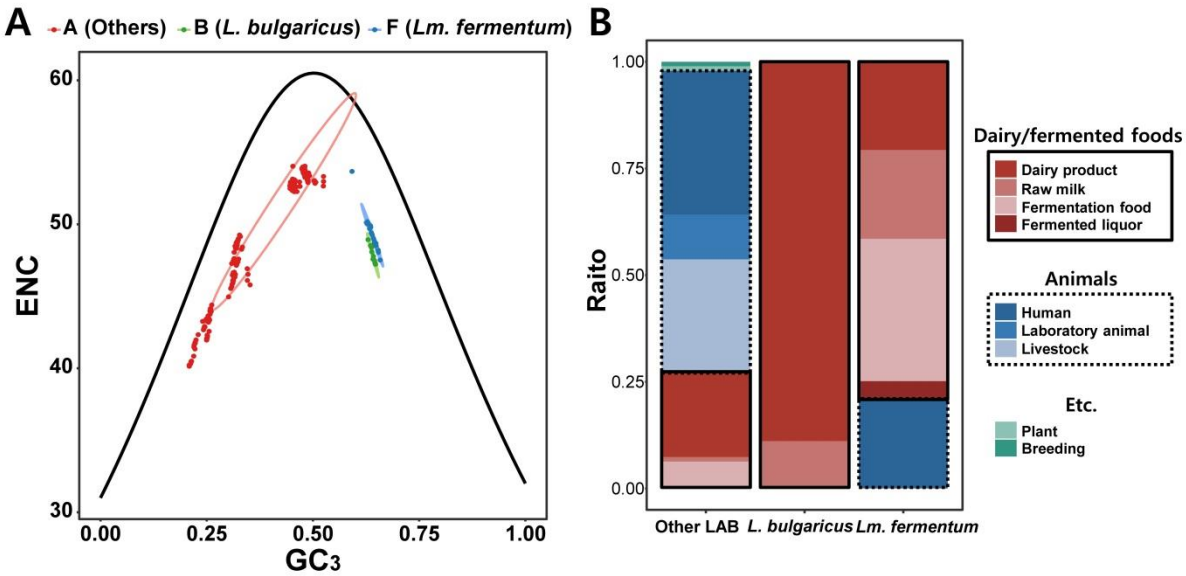


Figure 3-6. ENC-GC3 plot and isolation source of lactic acid bacteria

(A) Blue circles and green circles denote *Lm. fermentum* and *L. bulgaricus*, respectively, while red circles indicate the other species. The thick black curve is the expected curve derived from the positions of strains when the codon usage is determined only by the GC3s composition. **(B)** All isolation sources were divided by main categories: Dairy/fermented foods (red-colored and black lined bars), Animals (blue-colored and dot-lined bars), etc. (green-colored bars).

3.4.5. Detection of a candidate gene related to elevated GC content and classification of isolation source

Gene comparisons within the genus were performed to find gene factors related to changes in GC content. In the genus *Lactobacillus*, 98 genes were confirmed to significantly differ in copy number from *L. bulgaricus* and other *Lactobacillus* genus bacteria (Figure S6). More than half of the genes were classified in the metabolism category in protein functional classification. Following that, it was confirmed that 22.1% of genes were classified in the information storage and processing category. Comparing the species belonging to *Limosilactobacillus* with *Lm. fermentum*, 41 genes were confirmed to be different from the two groups. In COG annotation, the genes were most frequently in the Information storage and processing category (57.5%), followed by metabolism (34.3%).

The original source of bacterial strains was compared to investigate the relationship between GC content and microbial habitat (Figure 6B). All strains of *L. bulgaricus* were isolated from dairy/fermented food, and 87.5% of them were separated from dairy products. In *Lm. fermentum*, diary/fermented foods were the main isolation source as 79.17%, followed by animal hosts. All strains from animals were isolated from humans. On the other hand, other LAB had various isolation sources, and 70.53% of other LAB strains were separated from the animal hosts.

3.5. Discussion

In this study, two species (*L. bulgaricus* and *Lm. fermentum*) with high GC content for the genome size were identified among lactic acid bacteria. GC content is understood as a result of bacterial evolution, but most previous studies on the relationship between GC content and evolution have been conducted over the bacterial kingdom (Wu, Zhang et al. 2012). Therefore, we conducted this study to more specifically understand genetic characteristics with a high GC content of lactic acid bacteria. In all kingdoms including Bacteria, thymine occurrences due to cytosine deamination are frequently observed, and this nucleotide transition leads to loss of GC pairs (Lind and Andersson 2008). In addition, low GC content is often found to reduce replication costs in bacteria exposed to nutrient-limited and malnutrition environments (Kogay, Wolf et al. 2020). Therefore, the two species with high GC contents in *Lactobacillaceae* are unusual. In order to understand bacterial evolution, many studies have been conducted on the relationship between various genetic factors and GC content. Co-evolution of codon usage with tRNA for translational optimization, transposable elements associated with HGT, and DNA polymerase III (α subunit) used for replication were nominated as related factors (Zhao, Zhang et al. 2007, Higgs, Ran et al. 2008, Acman, van Dorp et al. 2020). A comparison of the genetic factors correlated with the GC content was also performed, but no difference but the number of RNAs was found. These

results suggest the elevated GC content of *L. bulgaricus* and *Lm. fermentum* may be a different factor from the suggested in previous studies. Comparison of genetic factors presented in previous studies limits the scope of inferences about factors affecting GC content.

By comparing the density of GC content by gene and GC ratio by codon position, it was verified that the elevated GC content in *L. bulgaricus* and *Lm. fermentum* showed a significant increase in CDS, especially GC3. This change was also presented in the previous study on *L. bulgaricus*, and this study also showed similar results (van de Guchte, Penaud et al. 2006). However, not only GC3 but also GC1 and GC2 showed higher GC contents than other LAB, indicating that the increased GC content was not limited to GC3. Also, the increased GC pair at each nucleotide position and the probability of non-synonymous mutation when mutation per codon occurs is proportional (Table 2). This GC pair enhancement makes it possible to infer that the high GC mutations found in both species occurred within a range that minimized changes in amino acids. The constant sum of codons encoding the same amino acid but different third nucleotides indicates that most of the elevated GC content is caused by a synonymous mutation. Since non-synonymous mutations in CDS are limited because they cause functional changes, synonymous mutation better shows the result of adaptation to the environment (Parmley and Hurst 2007). In the comparison of codon usage, it was confirmed that codons with high GC% were used

frequently in *L. bulgaricus* and *Lm. fermentum*. In particular, the RSCU analysis suggested that the preferred third nucleotide in the two species was cytosine. These differences were agreed upon in the overall CDS comparison and orthologous gene comparison. Interestingly, this change in codon usage was not found in the stop codon. Although stop codon is possible to synonymous mutation, codon preferences were similar. This observation suggests that the factor due to the codon usage difference is after the transcription, which does not affect the stop codon. ENC-GC3 plot and Kullback-Leibler divergence were used to compare the degree of codon bias and the information entropy for codon use in each group. In both analyzes, it was confirmed that *L. bulgaricus* and *Lm. fermentum* had a less biased codon and amino acid usage than other LAB through low D_{KL} and high ENC values. This indicates that the amino acid and codon usage of the two species are similar to an ideal random distribution than other LAB. At the same time, through comparison with the expected ENC-GC3 curve, it was confirmed that the codon bias in each group was due to selection, not mutation. However, further analysis is needed to understand what selection pressure was given to each group.

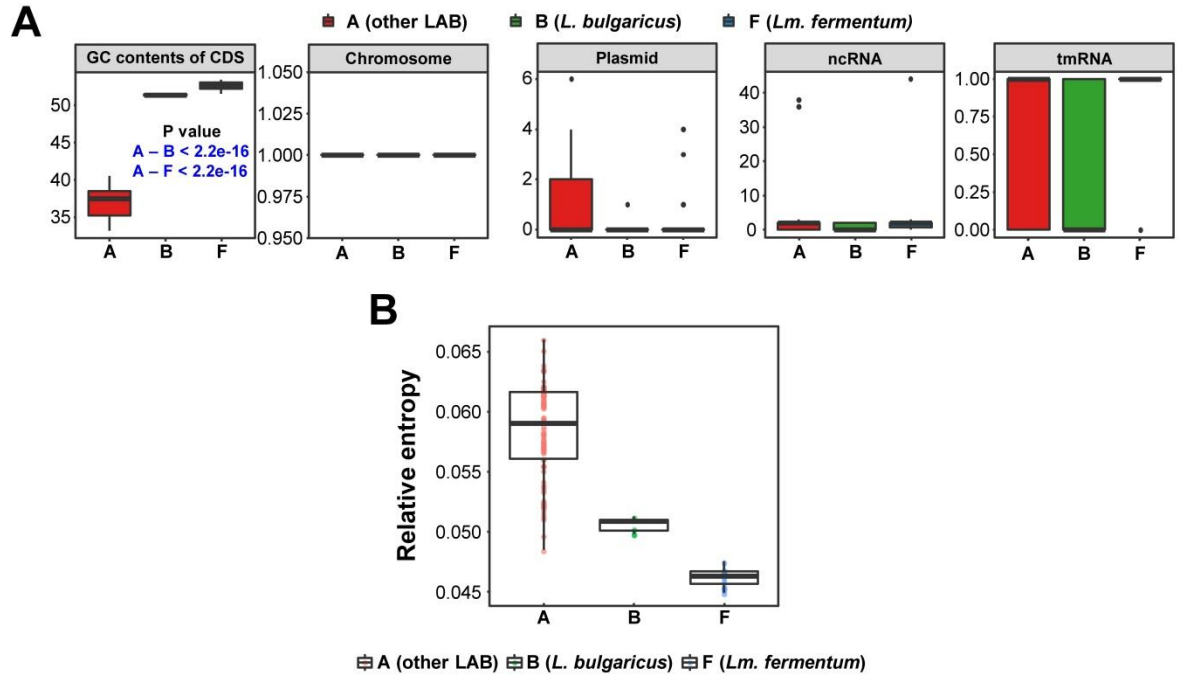
There are two main assumptions for understanding the increased GC contents in Bacteria. The first is a strategy to increase the efficiency of the bacterial organism by changing GC contents. In general, synthesizing guanine and cytosine in the biosynthesis of nucleotides requires more

energy than synthesizing adenine and thymine (Rocha and Danchin 2002). Therefore, a genome with high GC content might be quite inefficient in terms of energy. However, in amino acid synthesis, one GC-rich codon encodes an amino acid with low synthesizing energy in many cases (Du, Zhang et al. 2018). In addition, energy is more expensive and diverse in the process of synthesizing amino acids than in the process of synthesizing RNA. The energy required to generate two nucleotides is 4.6 ATP on average, whereas 13.2 ATP is required for amino acids, which is relatively higher (Chen, Lu et al. 2016). In addition, each mRNA template is translated multiple times, thus amplifying the selection for amino acid use. For example, in *Escherichia coli*, 47.3% of the production capacity is consumed for amino acids, but only 11.3% for RNA and nucleotides (Chen, Lu et al. 2016). Thus, encoding a cheaper amino acid is more efficient than favoring an expensive nucleotide, and the overall energy efficiency in the cell is improved by increasing the GC content of the genome. However, no change in amino acid usage pattern was detected with a significantly increased GC3 in our results. The second assumption is that it is the result of an adaptation of rapid reproduction in a nutrient-rich environment. Bacterial GC content is associated with environmental niches and their lifestyle. It is well known that GC content is changed by the result of horizontal gene transfer, and obligate and non-free-living organisms have GC-poor genomes. Also, bacteria evolve toward having low GC content in nutrient-limited and

competitive environments (Mann and Chen 2010). Therefore, a bacterium in a nutrient-rich environment with few competitors would have a high GC content within a range that did not alter amino acids. Besides, a study demonstrated that codon optimization is associated with improved protein production by increasing GC content by not adjusting translation in 2016 (Newman, Young et al. 2016). It indicates that protein production can be increased through mutation to guanine or cytosine. Most of *L. bulgaricus* and *Lm. fermentum* was isolated from dairy and fermented foods, which are a nutrient-rich environment with few other bacteria. It could allow fast protein production and frequent generation. Thus, genetic characteristics of *L. bulgaricus* and *Lm. fermentum* (increase of synonymous mutation, less biased codon and amino acid use patterns, codon usage that might be affected after transcription, and more efficient codons and amino acids use) seems to be an adaptation to a nutrient-rich environment. The rapid growth rates of both species uphold this assumption (Chervaux, Ehrlich et al. 2000, Rezvani, Ardestani et al. 2017). Likewise, the result of the comparison of functional protein genes can be interpreted in the same context. In the result, the genes related to replication, transcription, and amino acid metabolism were mainly found. In detail, genes related to purine and pyrimidine synthesis and amino acid metabolism were discovered, implying active nucleotide and amino acid metabolism in *L. bulgaricus* and *Lm. fermentum*.

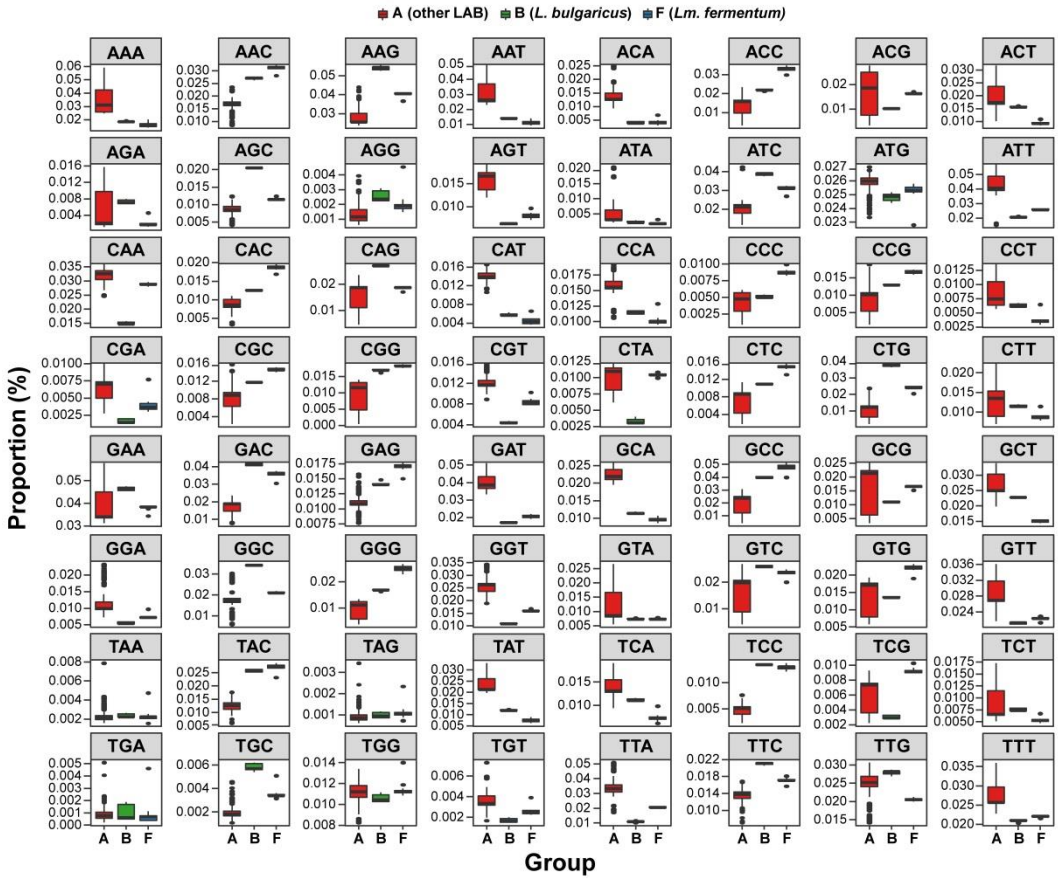
Therefore, we suggest the genetic characteristics of *L. bulgaricus* and *Lm. fermentum* is a result of an environmental adaptation.

Our study presented species with high GC content in lactic acid bacteria through genome comparison analysis, and genetic characteristics in species with high GC content were described. We identified that the elevated GC content was affected by CDS, especially GC3, increasing synonymous mutations. Although there was a limitation to provide a direct factor increasing the GC content, it was suggested that more diverse codons and amino acids were used in lactic acid bacteria with high GC content through codon usage comparison and statistical analysis. We hope that our results will provide a detailed understanding of the GC content and bacterial evolution of lactic acid bacteria.



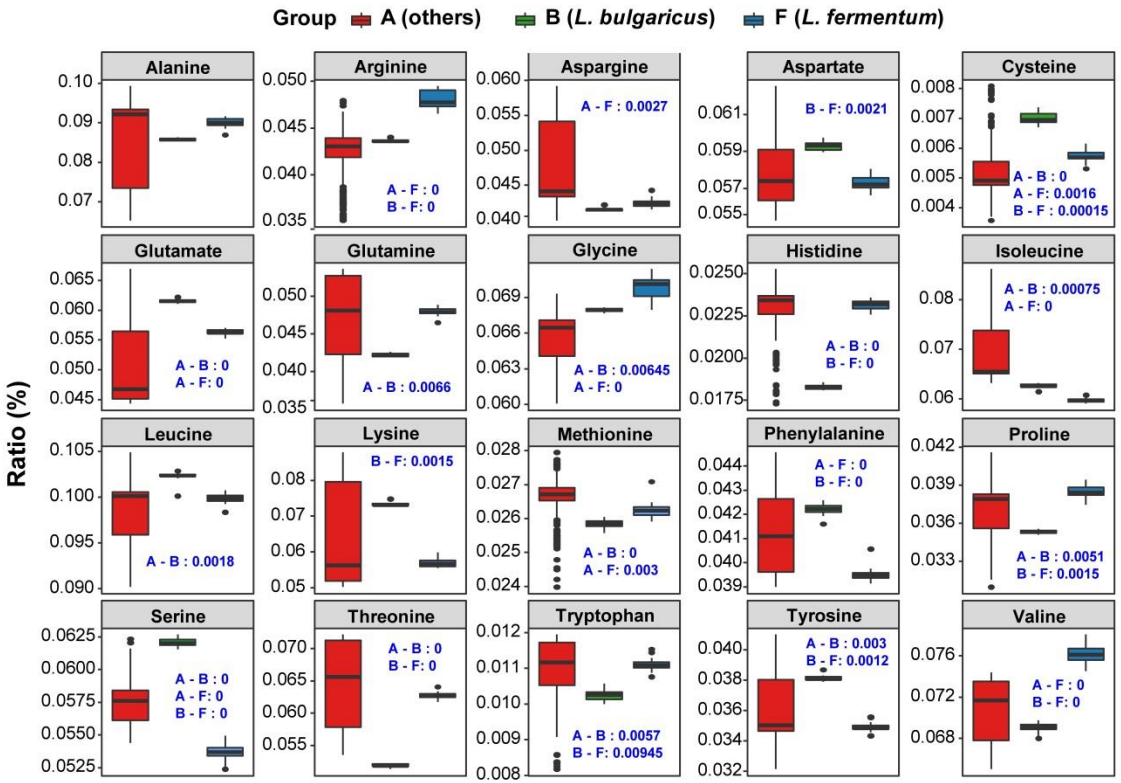
Additional file 4 - Figure S3-1. Comparison of potential genetic factors related to GC contents

(A) Genetic content comparison of lactic acid bacteria with 1.5-2.5Mb of genome size. A significant difference between comparison groups is denoted by blue-colored bold text with P-value limited under 2.2e-16. **(B)** Kullback–Leibler divergence (D_{KL}) of comparison groups for amino acid use. The value was plotted using a box plot, and each value was showed a point.



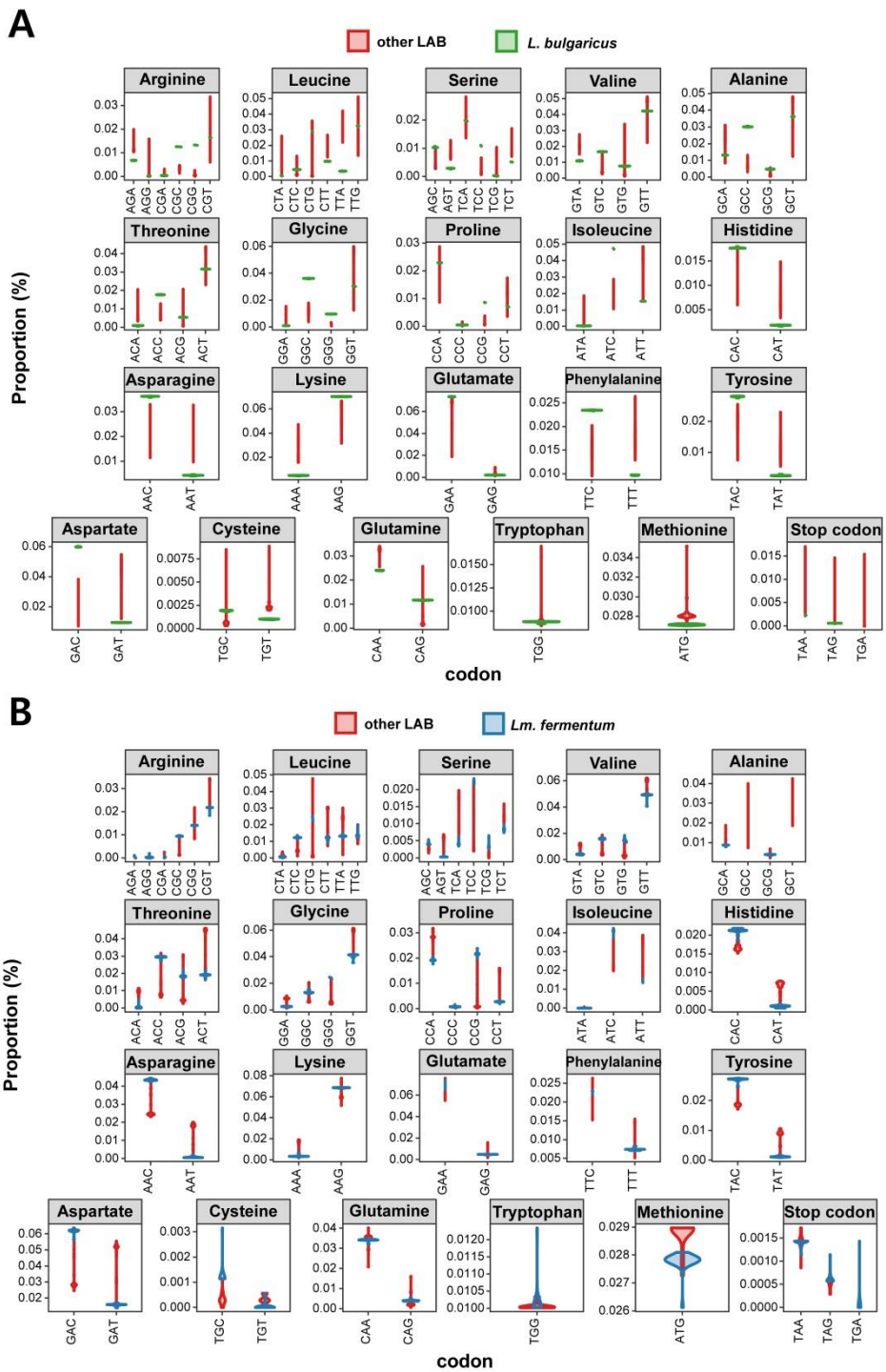
Additional file 5 - Figure S3-2. Comparison of codon usage

The x-axis shows three comparison groups, and the y-axis represents codon proportion, which was calculated as codon frequency dividing a total number of codons.



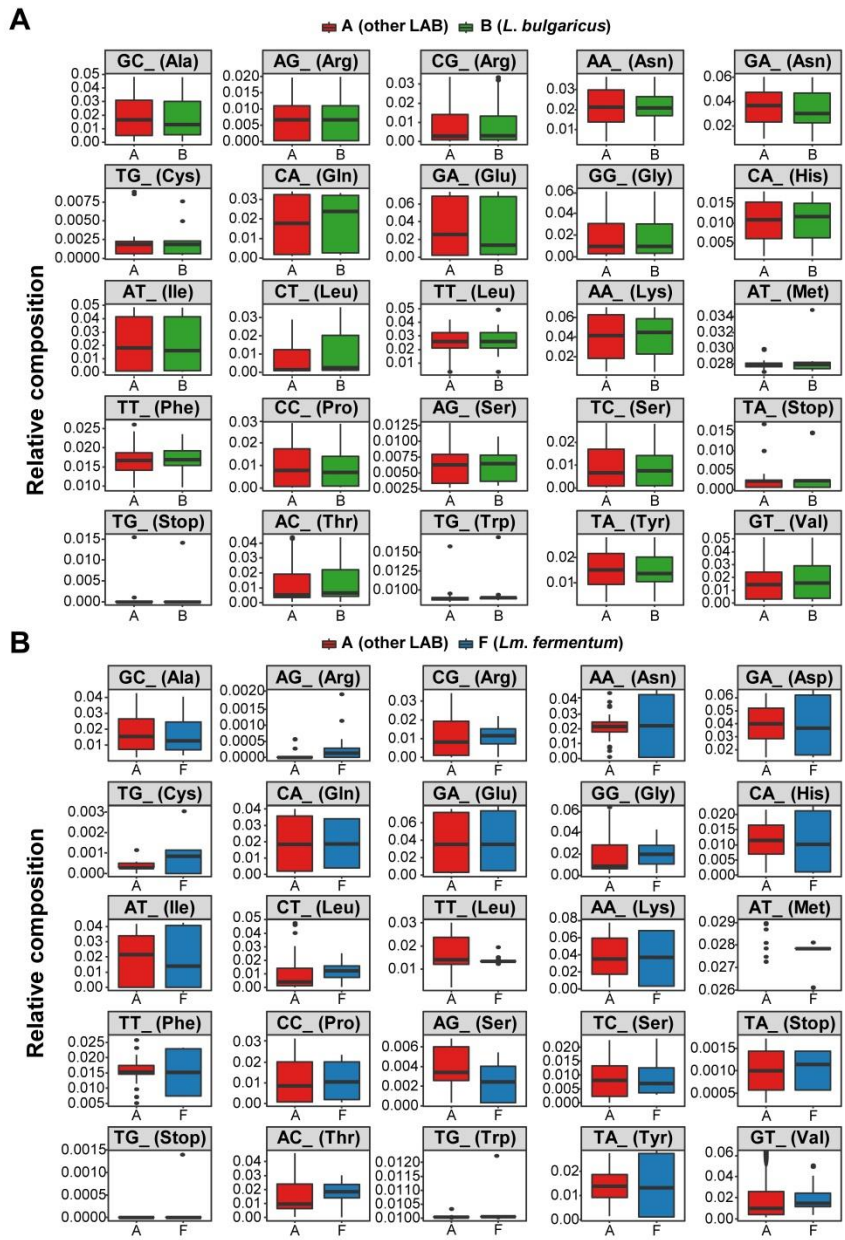
Additional file 6 - Figure S3-1. Comparison of amino acid preference

The ratio of the y-axis shows a percentage of total used amino acids in coding sequences. A significant difference between comparison groups is denoted by blue-colored bold text, and the P-value under 2.2e-16 is marked as 0.



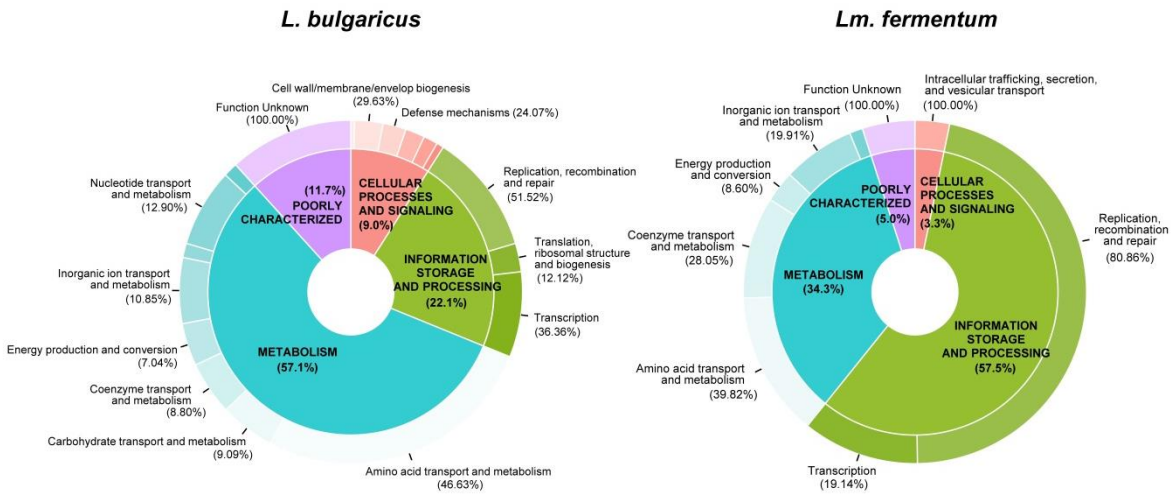
Additional file 7 - Figure S3-2. Violin plot for codon usage of *L. bulgaricus* and *Lm. fermentum* comparing orthologous genes

The x-axis shows codons encoding the same amino acid, and the y-axis represents codon proportion, which was calculated as codon frequency dividing the total number of codons in coding sequences.



Additional file 8 - Figure S3-3. Proportion comparison of the sum of non-synonymous codons

The x-axis shows comparison groups, and the y-axis represents the codon ratio obtained by dividing the sum of non-synonymous codons (accurately, the sum of the codons that encode the same amino acid and the third nucleotide is different) by the total number of codons.



Additional file 9 - Figure S3-4. Functional classification of protein coding genes

Genes that had significant copies comparing species included in the same genus were annotated and classified by function categories. The inner circle presents main categories, and the outer circle shows subcategories.

This chapter was published in *Frontiers in Microbiology* (2021)
as a partial fulfillment of Soomin Jeon's Ph.D program.

Chapter 4. Complete Genome Sequence of the Newly Developed *Lactobacillus acidophilus* Strain With Improved Thermal Adaptability

4.1. Abstract

Lactobacillus acidophilus (*L. acidophilus*) is a representative probiotics and is widely used in many industrial products for its beneficial effects on human and animal health. This bacterium exposes to harsh environments such as high temperatures for manufacturing industrial products, but cell yield under high temperatures is relatively low. To resolve this issue, we developed a new *L. acidophilus* strain with improved heat resistance while retaining the existing beneficial properties through the adaptive laboratory evolution (ALE) method. The newly developed strain, *L. acidophilus* EG008, has improved the existing limit of thermal resistance from 65˚C to 75˚C. Furthermore, we performed whole genome sequencing and comparative genome analysis of wild-type and EG008 strains to unravel the molecular mechanism of improved heat resistance. Interestingly, only two single nucleotide polymorphisms (SNPs) were different compared to the *L. acidophilus* wild-type. We identified that one of these SNPs is a non-synonymous SNP capable of altering the structure of MurD protein through the 435^{th} amino-acid change from serine to threonine. We believe that these results will directly contribute to an industrial field where *L. acidophilus* is applied. In addition, these results make a step forward in understanding the molecular mechanisms of lactic acid bacteria evolution under extreme conditions.

Keywords

Adaptive laboratory evolution, *Lactobacillus acidophilus,* heat resistance, whole genome sequencing, Bacterial evolution

4.2. Introduction

Lactic acid bacteria (LAB) is a gram-positive bacteria which produces lactic acid as a fermentation product (Makarova, Slesarev et al. 2006, Zheng, Wittouck et al. 2020). Since LAB has been mainly applied in dairy products or fermented foods for humans and animals, it has been domesticated to satisfy profitable features (Steensels, Gallone et al. 2019). For example, these bacteria are highly resistant to acids and bile salts, making them widely used for industrial purposes such as food manufacturing (Menconi, Kallapura et al. 2014). Functionally, they play a beneficial role in inhibiting the growth of pathogens by producing antimicrobial compounds such as lactic acid, hydrogen peroxide, and bacteriocin (Mohankumar and Murugalatha 2011).

Lactobacillus acidophilus (*L. acidophilus*) is a representative LAB species that has been well studied in its physiology and functionality. The beneficial health effects of *L. acidophilus* have been shown in studies of various diseases such as innate immunity (Klein, Friedrich et al. 2008, Foysal, Fotedar et al. 2020), inflammatory bowel disease (Peran, Camuesco et al. 2007, Park, Choi et al. 2018), and colon cancer (Zhuo, Yu et al. 2019). In addition, as scientific facts about various effects such as skin wrinkle improvement (Chahuki, Aminzadeh et al. 2019), skin moisturizing (Im, Lee et al. 2018), and vaginal cleansing (Bertuccini, Russo et al. 2017) are revealed, the scope of utilizing *L. acidophilus* in the industrial field is

gradually expanding. Based on the scientific evidence of these health benefits, *L. acidophilus* holds an important position in the probiotic market, and a variety of commercial strains have been discovered such as *L. acidophilus* NCFM (Altermann, Russell et al. 2005), *L. acidophilus* LA-1 and LA-5 (Schillinger, Yousif et al. 2003, Matijašić, Obermajer et al. 2016), and *L. acidophilus* DDS-1 (Dash 2004). *L. acidophilus* is known to have the same high acid, bile salt, and osmotic resistance as the common LAB (Hutkins, Ellefson et al. 1987, Chou and Weimer 1999).

When LAB are used for manufacturing industrial products, the bacterial strain is often exposed to extreme environments. Among various environmental factors, high temperature has a major influence on bacterial survival. For example, LAB are used as an animal feed additive, and they are usually manufactured in pellet form. Pellets are made by compressing ground feed and supplements including LAB by applying air above 80 ℃ (Skoch, Behnke et al. 1981). Some feed mills have compression temperatures that can reach 90 ℃ to destroy feed-borne pathogens such as *Salmonella* (Jones and Richardson 2004). After this process, the bacteria are useful to animals only if it survives in the pellet. Thermal and mechanical treatments have physiological and biological effects on living cells, such as denaturing proteins and altering enzymatic activity (Belhadj Slimen, Najar et al. 2016). It is appeared by a decrease in the cell viability. Therefore, strains such as *Saccharomyces cerevisiae* and *Bacillus subtilis* with high

survival rates under heat treatment are mainly used (Haldar, Ghosh et al. 2011, Nguyen, Nguyen et al. 2015). Among the lactic acid strains, *Enterococcus faecium* is mainly utilized (Boney, Jaczynski et al. 2018). *L. acidophilus* has been suggested as an antibiotic alternative by improving growth performance and nutrient utilization in the animal intestines, but their low thermal stability is a limitation for using as feed additives (Simon, Vahjen et al. 2005, Lan, Koo et al. 2017). Furthermore, in addition to direct heat treatment to LAB, a thermostable bacterial strain can help protect against heat-induced death or accidental thermal management defects during fermentation, thereby increasing cooling cost effectiveness (Matsushita, Azuma et al. 2016). Therefore, it is expected that improving the heat resistance of *L. acidophilus* will not only increase the industrial utility but also contribute to the expansion of the application range.

Encapsulation and heat pretreatment for heat-shock protein expression have been mainly studied as the methods for improving heat stress resistance of diverse bacterial strains, but these methods are not cost effective (Xu, Gagné-Bourque et al. 2016, Chen, Tang et al. 2017). On the other hand, adaptive laboratory evolution (ALE) artificially stimulates natural evolution in a laboratory setting, making it relatively easy to improve the desired phenotype of targeted strain (Portnoy, Bezdan et al. 2011, Dragosits and Mattanovich 2013). Research to increase the heat resistance of bacteria by applying the ALE method has been conducted in

various species such as *Escherichia coli* (Riehle, Bennett et al. 2003, Rudolph, Gebendorfer et al. 2010) and *E. faecium* (Min, Yoo et al. 2020). Likewise, studies to improve the thermal resistance of LAB have been steadily carried out over the past 20 years, but most of the studies have only confirmed the improved cell viability below 65℃. For example, survival rates of *Lacticaseibacillus paracasei* DPC1919, DPC2102, and DPC 2013 strains were evaluated at 56 to 67.5℃ (Jordan and Cogan 1999). In other study, the thermal resistance of *L. acidophilus* LA1-1 were measured at 37 to 58℃ (Kim, Perl et al. 2001). There have also been attempts to develop *L. acidophilus* EG008 strains using ALE such as *L. acidophilus* NCFM at 65℃ (Kulkarni, Haq et al. 2018) and *Lactiplantibacillus plantarum* Lp 998 at 45 to 55℃ (Ferrando, Quiberoni et al. 2015).

Based on these rationales, the primary goal of this study is to develop a strain of *L. acidophilus* that can withstand conditions of 65℃ or higher through the ALE method. The secondary goal is to minimize changes in the genetic background to maintain the functional advantages of the existing *L. acidophilus* strain as much as possible. We confirmed this through complete genome analysis using long-read sequencing.

4.3. Materials and Methods

4.3.1. Strain identification and bacterial culture

Probiotic colony was isolated from a fermented dairy food. To identify bacterial species, 16S rRNA genes were sequenced by Macrogen Inc. (Seoul, Korea). Sequencing reactions were performed in the DNA Engine Tetrad 2 Peltier Thermal Cycler (BIO-RAD, Hercules, California) using the ABI BigDye (R) Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Beverly, Massachusetts). Primers used for single-pass sequencing were as follows: forward primer 27F (5'-AGAGTTTGATCCTGGCTCAG-3') and reverse primer 1492R (5'-GGTTACCTTGTTACGACTT-3'). To remove the unincorporated terminators and dNTPs, the fluorescent-labeled fragments were purified by the method that Applied Biosystems recommends. The samples were injected to electrophoresis in an ABI 3730xl DNA Analyzer (Applied Biosystems). 16S rRNA sequences were compared with the NCBI database using BLAST (Johnson, Zaretskaya et al. 2008). For bacterial culture, all strains were propagated statically in deMan Rogosa Sharpe broth (MRS broth; Difco Laboratories, Detroit, Michigan) or on MRS agar (1.5% [wt/vol]) under aerobic condition at 37℃ without shaking. Gram-staining was performed using the BD BBL™ Gram stain kit following the manufacturer's protocol. For viable cell counting, serially diluted cultures were poured into the MRS plates and inoculated at 37℃ for 48 hours. The viability of the cell was counted in the colony-forming unit per mL

(CFU/mL). Cell density was measured by the absorbance at 600nm (OD_{600}; Optical density spectrophotometrically measured at 600 nm wavelength) using an OPTIZEN POP UV-visible spectrophotometer (KLAB, Daejeon, Korea).

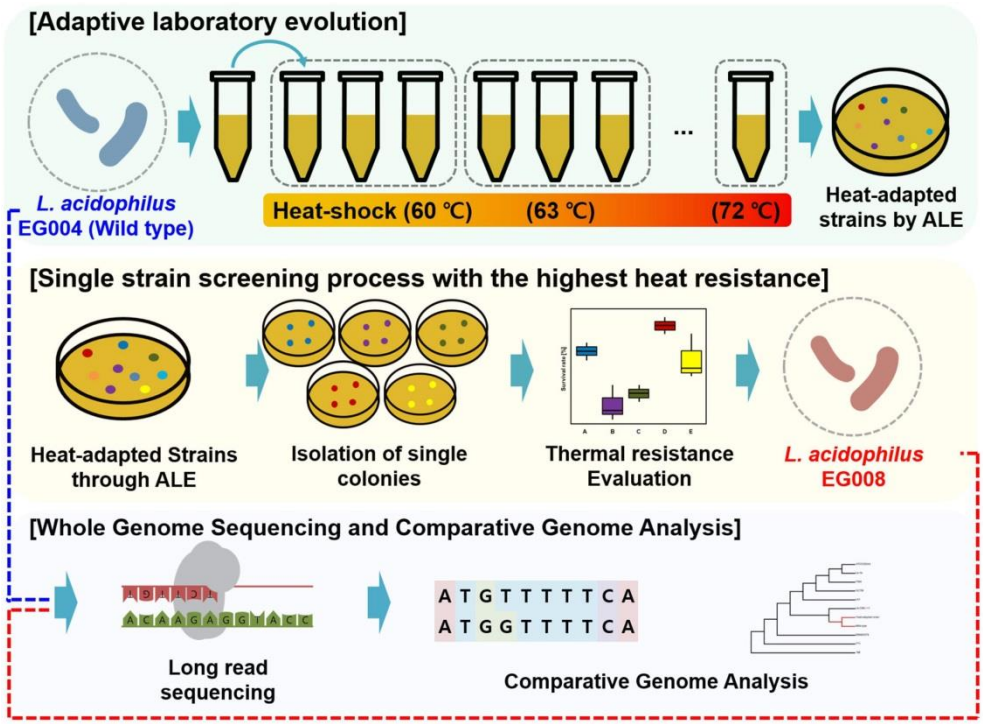


Figure 4-1. Schematic diagram of the study to develop *L. acidophilus* EG008 strain

This diagram shows the overall process for developing an EG008 strain of *L. acidophilus*. The first step is the thermal adaptation of *L. acidophilus* EG004 strains through adaptive laboratory evolution. The second step is the process of selecting the colonies that have the single strain with the highest heat resistance among the various strains that have been thermally adapted. The final step is the *in-silico* analysis step to identify and compare the whole genome sequence of *L. acidophilus* EG004 and EG008 strains, respectively.

4.3.2. Adaptive Laboratory Evolution and screening a thermal adapted strain

To induce newly strain improved thermal tolerance, ALE experiment at high temperature was applied. Development of Heat-adapted strain consisted of two steps: heat adaptation to the highest survival temperature and single strain selection with thermotolerance (Figure 1). *L. acidophilus* EG004 strain was used as wild-type and cultured for 16 hours was prepared for heat adaptation to the highest temperature. Ten microliters of cells were injected into 990 microliters of pre-heated MRS broth at 60˚C by heat block (ALB64; FINEPCR, Gyeonggi-do, Korea). Heat treatment was applied for 1 minute. After heat treatment, it was allowed to cool down at room temperature for 5 minutes, and cultivated at 37˚C for 24 hours. After two iterations, the identical process was repeated with increased temperature by 3˚C until incubation was impossible. All strains of each step were stored with 25% glycerol at -80˚C. Next, we performed a single strain selection from the microbial population obtained through the final ALE treatment (“A001F8-72”). It was performed because the final result of ALE is presumed to be a group of individuals with random mutations rather than a single individual. The evaluation temperature was determined as the temperature above the critical point of EG004 strain. Five different colonies were picked from A001F8-72. A hundred microliters of cells were injected in 900 microliters of pre-heated saline at 66˚C by heat block and cells were

heated for 1 minute. After cooled down, the diluted cells were plated on MRS agar to determine viable cell count. The cell survival ratio was expressed by dividing the viable cell count after heat stress by the initial viable cell count. The strain had the most improved resistance to high temperature was named as “EG008” strain, and used for the comparative analysis.

4.3.3. Assessment of phenotypical changes

In order to assess improved thermal capacity, cell viability at 55 - 75˚C was assessed. Cells in the mid-exponential phase were prepared to OD_{600} at 1.3, corresponding to approximately $1x10^9$ CFU/ml. A hundred microliters of cells were injected into 900 microliters of pre-heated saline by heat block. After heat was applied for 1 minute, cells were plated on MRS agar to quantify the viable cell count. Thermal resistance was presented as the survival ratio. To check other phenotypical changes induced by heat stress, acid and salinity tolerance were measured. Inoculated cells were harvested by centrifugation at 12,000 rpm for 10 min at 4℃. After the supernatant was removed, the pellet was washed 2 times with 1X phosphate buffer (1X PBS). It was resuspended into 1 ml of PBS buffer adjusted to acidified solutions (up to pH 2.0 and 7.0) or salted solutions (7.5% and 15.0%). Cells were exposed to the acidic solution for 2.5 hours and the high salt solution for 3.5

hours at 37℃. The viable cell count was measured in order to assess cell survival, and it was calculated by the same method as thermal resistance.

4.3.4. Bacterial kinetics

The biomass was measured by dried cell weight. Cultured bacteria in 1000ml of MRS broth were centrifuged at 4000rpm for 10 minutes, washing twice, and then drying them at 60℃ for 3 days. All measurements were repeated three times. The concentration of glucose was measured by HPLC in the culture solution filtered through a 0.45um membrane filter (HPLC machine: Dionex ultimate 3000 (Thermo Dionex, USA / pump, autosampler, oven), Detector: Shodex RI-101(Shodex, japan), and Column: Sugar-pak (Waters, 300*6.5mm, USA)). Glucose (Junsei chem, 98%) was used as glucose standard. Since batch culture was performed in an Erlenmeyer flask, the Monod equation was used to calculate the factors (Okpokwasili and Nweke 2006).

4.3.5. Statistical analysis

All experiences were performed with 3 replications to check if there is any experimental bias except single strain selection from A001F8-72. The student’s t-test was used to compare assessments of thermal, acidic, and salinity tolerance. We considered a 5% significance level. The FDR correction method was used for multiple testing issue.

4.3.6. Whole genome sequencing

Whole genome sequencing was served by Macrogen Inc. using SMRT Sequencing. Samples were prepared according to a guide for preparing SMRTbell template for PacBio sequencing. NanoDrop spectrophotometer (Thermo Scientific, Waltham, Massachusetts) and PicoGreen quantified the concentration of gDNA. All samples passed screening QC criteria. For PacBio Sequel sequencing, 5ug of gDNA was served for 10 kb library preparation. For gDNA less than 17 kb, the actual size distribution was evaluated by Bioanalyzer 2100 (Agilent)d. Sheared gDNA using g-TUBE (Covaris Inc., Woburn, Massachusetts) was purified by AMPurePB magnetic beads (Beckman Coulter Inc., Brea, California) if the apparent size was greater than 40 kb. A total of 10uL library was arranged using PacBio DNA Template Prep Kit 1.0. SMRTbell templates were annealed using Sequel Binding and Internal Ctrl Kit 3.0. The Sequel Sequencing Kit 3.0 and SMRT cells 1M v3 Tray was conducted for sequencing. The PacBio Sequel platform captured SMRT cells (Pacific Biosciences, Menlo Park, California). The next steps were followed as the PacBio Sample Net-Shared Protocol. Raw data from PacBio RS II was assembled by PacBio SMRT portal system and HGAP4 tool assembled. Genome assembly was conducted with genome size parameter was set to 3 Mb. Assembled contig with low quality such as a short length (<20,000 bp) and low coverage

(<50X) was eliminated. To correct assemble errors in the assembled genome sequence, polishing process was repetitively conducted with quiver algorithm until genomic variants were not found. Assembled genome was circularized by Circlator (Hunt, De Silva et al. 2015).

4.3.7. Annotation of genomic information

The genomes of *L. acidophilus* EG004 and EG008 strains were annotated and genes were categorized by protein functions using Rapid Annotations using Subsystems Technology (RAST) server with version 2.0 (Aziz, Bartels et al. 2008). To identify its functionality and safety as a probiotics, several genetic factors were detected. Antibiotic-resistant gene was inspected by NCBI BLAST with ARG-ANNOT and CARD databases (Gupta, Padmanabhan et al. 2014, Jia, Raphenya et al. 2016) Virulence factor and prophage gene were identified by VirulenceFinder 2.0 (Joensen, Scheutz et al. 2014) and PHASTER (Arndt, Grant et al. 2016), respectively. IslandView4 discovered genomic island (Bertelli, Laird et al. 2017). Identification of bacteriocin was carried out using BAGEL4 (van Heel, de Jong et al. 2018). To detect variants in EG008 strain, a comparative analysis was conducted. To find a singleton in each strain, orthologous gene and singleton definition were conducted with OrthoVenn2 software (Xu, Dong et al. 2019). Identified singletons were double-checked by the NCBI BLAST.

Single nucleotide variants were detected by nucmer including MUMmer 3.23 (Kurtz, Phillippy et al. 2004).

4.3.8. Comparative genomic analysis with Lactobacilaceae family

To identify closeness between our strains and Lactobacillaceae family, 22 genomes were used and compared with EG004 and EG008 strains. 22 genomes of Lactobacilaceae and a genome of *Bacillus subtilis* were downloaded from the NCBI website. The genome of *B. subtilis* was used as an outgroup to make phylogenetic root of Lactobacillaceae family. Average Nucleotide Identity (ANI) was calculated by JSpecies1.2.1 (Richter and Rosselló-Móra 2009). 16S rRNA sequences were investigated by RNAmmer and aligned by ClustalW2.1 (Lagesen, Hallin et al. 2007, Larkin, Blackshields et al. 2007). The phylogenetic trees were generated by MEGA X with bootstrap 1,000 using the Neighbor-joining method (Kumar, Stecher et al. 2018). To compare functional gene contents, protein prediction of 22 Lactobacillaceae genomes was performed by the RAST server. We certified if detected variants in EG008 strain exist in other Lactobacillaceae. In the case of genic SNP, the gene including SNP was used as a blast query for comparison. In the case of intergenic SNP, a sequence with length of 301 base pairs including the variant was used. Multiple sequence alignment by ClustalW2.1 was used to confirm sequence detection. In comparison of singleton, BLAST was utilized and Identification of gene location was

employed by Artemis (Rutherford, Parkhill et al. 2000). The pan-genome analysis was conducted using PGAP program with the cut-off filtering of E-value (<1e-10) (Zhao, Wu et al. 2012).

4.4. Results

4.4.1. Development of heat-resistant *L. acidophilus* strain based on ALE method

We isolated *L. acidophilus* EG004 strain from fermented dairy food and identified using 16S rRNA sequencing (Figure S1). First, in an experiment to find out the limit of heat resistance of the wild-type strain, we found out that it was unculturable at 75˚C for 99 generations. To overcome this limit of heat resistance, we developed the EG008 *L. acidophilus* strain by applying ALE with different temperatures from 60˚C to 72˚C (Figure 1). After the application of the ALE method, preliminary heat screening was conducted to select strains with the identical genotype that can significantly improve heat resistance in the wild-type population. As a result, *L. acidophilus* cultured in EG008 colony showed the highest possible improvement in heat resistance (47.768%) than others (Table 1).

Table 3-1. Results of preliminary investigations on heat resistance improvement for a single strain of wild type *L. acidophilus* A001F8-72

Colony	Before (CFU/ml)	After Heatshock (CFU/ml)	Survival Rate (%)	SD
A001F8-72-1	4.10e+08	1.67e+08	40.899%	0.036
A001F8-72-2	4.57e+08	1.13e+08	25.213%	0.065
A001F8-72-3	5.17e+08	1.47e+08	28.855%	0.034
A001F8-72-4	3.47e+08	1.65e+08	47.768%	0.014
A001F8-72-5	4.67e+08	1.81e+08	39.413%	0.058

The survival rate (%) was calculated by repeating the preliminary heat resistance improvement experiment 3 times from each colony containing a single strain. The evaluation was conducted at 66℃, where the survival rate rapidly decreased at.

4.4.2. Overcoming the limit of thermal resistance of *L. acidophilus* EG004 strain

As a result of conducting the experiment to determine the limit of heat resistance of the *L. acidophilus* EG004 strain, a significant decrease in survival rate was observed at 66˚C (Figure 2A). We hypothesized that the *L. acidophilus* EG008 strain developed by the ALE method would have a significantly improved survival rate at such thermal limitation. In the case of the newly developed *L. acidophilus* EG008 strain, the survival rate decreased as the temperature exceeded 66˚C, but a statistically significant improvement in the survival rate was observed compared to the EG004 strain at a 5% significance level (Figure 2B). In particular, at 66℃, the identified limit of heat resistance of *L. acidophilus* EG004 strain, an average of 39.11% improvement in survival rate was observed (P: 0.029). At temperatures above 68℃, the tendency to decrease with single-digit survival rate was similar to that of EG004, but a statistically significant improvement pattern of heat resistance was confirmed.

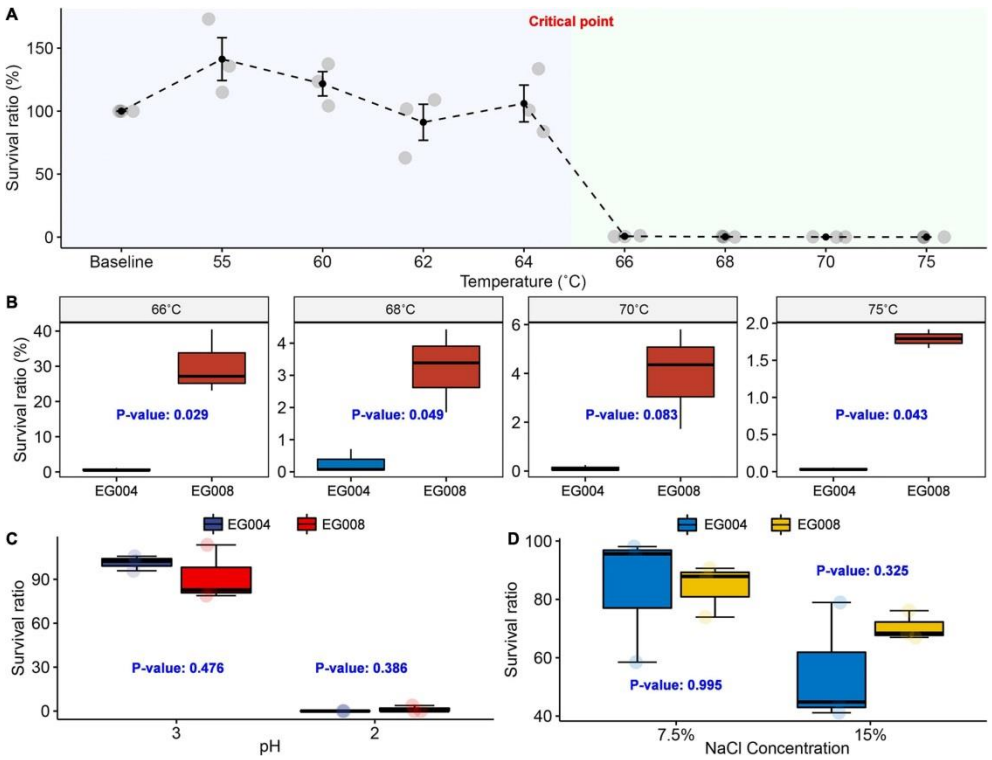


Figure 4-2. Experiment to confirm improved thermal resistance at critical temperature above 66 ˚C

(A) Survival curve of *L. acidophilus* EG004 strain under different temperature conditions. The survival rate was calculated after 1-minute exposure at each condition, and three repeated trials were performed. At 66 degrees Celsius, a sharp drop in survival was observed. **(B)** The newly developed *L. acidophilus* EG008 strain showed an improved survival ratio compared to the EG004 at critical point above 66 degrees Celsius. P-value is the result of the two-group t-test. **(C)** Investigation of survival rate in acidic condition. **(D)** Measurement result of change in survival rate according to salt concentration.

In previous studies that have developed heat resistant strains in diverse species, it has been reported that the developed heat resistant strain is additionally endowed with resistance to other types of stress such as acid and high salt concentration simultaneously (Min, Yoo et al. 2020). Based on this fact, we expected that the newly developed *L. acidophilus* EG008 strain, which was endowed with improved thermal resistance, could also have cross resistance. However, no statistically significant improvement in survival rate was observed under strong acidic conditions (Figure 2C) and salt concentration (Figure 2D), which suggests that genes related to heat resistance could be independent of genes related to acid and salt resistance in *L. acidophilus*. In addition, in order to verify whether the probiotic functionality and fermentation performance were maintained, a comparative evaluation was performed with *Lacticaseibacillus rhamnosus* GG (LGG) with well-proven functionality. As a result, the two experimental strains showed higher acid resistance compared to the LGG stain, and the salt resistance and bile salt resistance were similar (Figure S4). In assessment of fermentation performance, all estimated summary statistic of both strains were similar (Table S2).

4.4.3. Complete genomic analysis for *L. acidophilus* EG004 and EG008 strains

We completed the whole genome sequences from EG004 and EG008 strains through long-read sequencing technology (Table S1) to further reveal genomic characteristics to confirm that the newly developed strain is *L. acidophilus*. Constructed phylogenetic tree based on the 16S rRNA from 22 publicly available *Lactobacillaceae* whole genome sequences confirmed that the developed strains were one of the *L. acidophilus* species (Figure 3A). Of a total of 24 *Lactobacillaceae* familia, full-length sequences of 10 *L. acidophilus* strains used in the analysis showed a very high degree of similarity to each other (Avg. 99.835%) and relatively low similarity (Avg. 93.992%) to other *Lactobacillaceae* familia (Figure 3B). This result confirms once again that the heat resistant strain we developed is *L. acidophilus* and suggests that strains of *L. acidophilus* species have a specific genetic background in common, that is distinct from other *Lactobacillaceae* familia. Likewise, in the functional comparison based on gene annotation, no significant differences were observed between *L. acidophilus* strains (Figure 3C), but statistically significant differences were found in nine SEED terms at 5% significance level (Figure 3D) between *L. acidophilus* and other *Lactobacillaceae*. These results are ideal considering that our secondary goal is to maximize heat resistance while maintaining good properties such as antimicrobial properties of *L. acidophilus*. We confirmed that the genome of the *L. acidophilus* EG008 strain retained various antibacterial-related genes such as bacteriocin (Figure S2).

Furthermore, examining the features of the full-length genome, seven genomic islands and two prophage regions were found identically in both EG004 and EG008 strains (Figure S3). In addition, the pan and core genome analysis revealed that both strains are open pan-genome. Antibiotic resistance-related gene and a virulence factor were not found in both genomes. These results are one of the evidences showing that the newly developed EG008 strain has improved heat resistance, while maintaining the basic functional advantages of the existing wild-type *L. acidophilus*.

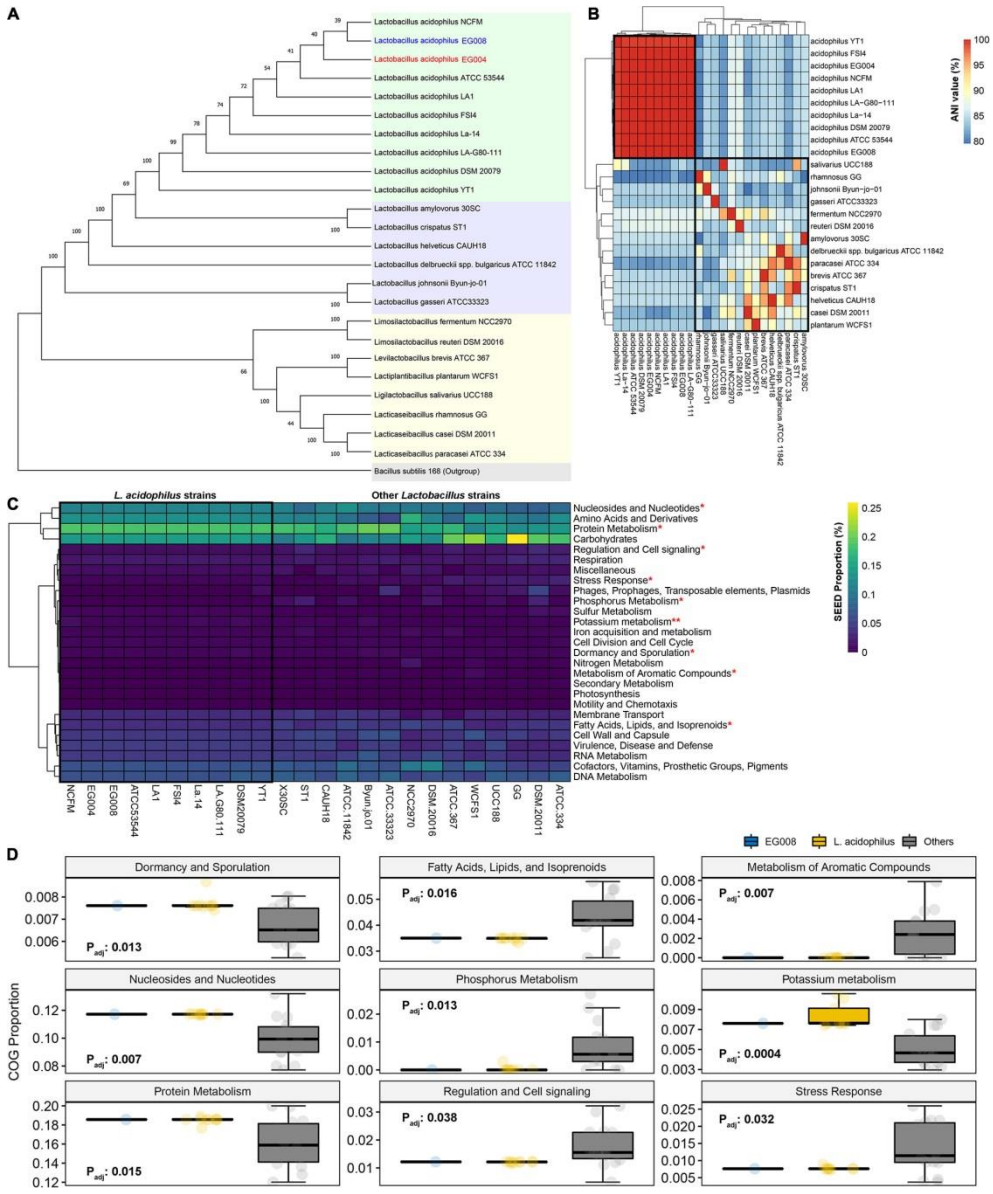


Figure 4-3. Comparative genome analysis for Lactobacillaceae based on the complete genome sequence of the newly developed *L. acidophilus* strain.

(A) Neighbor-joining tree based on 16S rRNA sequences for 24 Lactobacillaceae familia including the strain newly developed in this study. Bootstrapping was performed 1000 times, and the Bacillus subtilis 168 was considered as an outgroup. (B) Hierarchical

clustering results based on average nucleotide identity among 24 Lactobacillaceae familia. (C) Comparison of predictive SEED ratios between 24 Lactobacillaceae strains. (D) SEEDs with statistically significant differences between L. acidophilus and the rest of the Lactobacillaceae strains at 5% significance level after multiple testing adjustment.

4.4.4. Potential genetic factor related of heat resistance improvement through comparative genome analysis

We hypothesized that the genetic factors to improving heat resistance could be found by comparing the whole genome sequences of EG004 and EG008 strains. Interestingly, only two nucleotides were different in two genomes. As a result of performing multiple sequences alignment including other *L. acidophilus* strains whose full-length genome sequence were published, different genotype of these two SNPs was observed only in our newly developed EG008 strain. One of the two SNPs is located in the murD gene which synthesizes UDP-N-acetylmuramoyl-L-alanine--D-glutamate ligase (Figure 4A), while the other SNP is located intergenically between the galT gene (1,854,039 - 1,854,833bp) and the IdtD gene (1,854,986 - 1,856,194bp) (Figure 4B). Both mutations may be strong candidates for conferring heat resistance, but we further investigated for non-synonymous SNPs that are easy to be interpreted biologically. A variant in the murD gene of *L. acidophilus* EG008 strain causes substitution of nucleotide from T (Thymine) to A (Adenine). Subsequently, this substitution induces changes in amino acid from S (Serine) to T (Threonine) at 435^{th} amino acid residue of UPD-N-acetylmuramoyl-L-alanine-D-glutamate ligase (Figure 4C). Through protein structure analysis, we found that this substitution can trigger a change from hydroxyl to acyl group in the coil part of the protein (Figure 4D), which is located in the extracellular matrix (Figure 4E). These

results provide an insight into the mechanism of *L. acidophilus* strains with improved heat resistance at the molecular biology level.

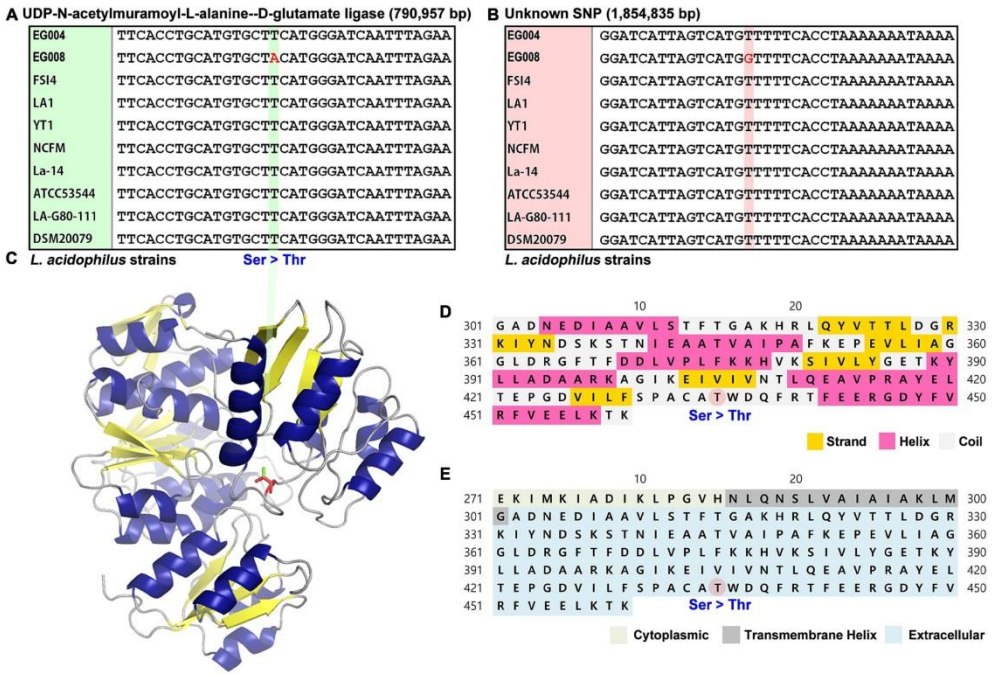


Figure 4-4. Two SNPs found by comparing the complete genome sequences of *L. acidophilus* EG004 and EG008 strains.

(A) A non-synonymous SNP is located at 790,957bp of genic region in EG008 strain. This genotype change causes the serine amino acid to be transformed into threonine. **(B)** A SNP found in the intergenic region of the EG008 strain. **(C)** Prediction of 3D structural change of MurD protein by the identified non-synonymous SNP. The serine to threonine changes triggers a change from hydroxyl to acyl group (Red) in the side-chain. **(D)** The 435th amino acid of the secondary structure is predicted to be the coli part by PSIPRED tool. **(E)** It is the part of the coil that connects the beta sheet of the third domain and the alpha helix, which is located on the extracellular matrix.

4.5. Discussion

In this study, we achieved the primary goal for improving the thermal resistance at 65℃ or more to secure the ease of industrial use of the *L. acidophilus* strain through stepwise ALE method (Figure 1). We have demonstrated that the newly developed *L. acidophilus* EG008 strain had better thermal resistance than wild-type at high temperatures of 65℃ to 75℃ (Figure 2B). Recently, studies on development of EG008 strains using ALE have been conducted (Kulkarni, Haq et al. 2018, Min, Yoo et al. 2020), but most of the experiments were performed at temperatures below 65℃ to ensure applicability in LTST pasteurization (Vaudagna, Sánchez et al. 2002). Therefore, the EG008 strain developed in this study is expected to be utilized for HTST pasteurization, which sterilizes at a relatively high temperature.

It is well known that repetitive heat stimulation employed in the ALE process can also induce resistance to other stresses such as acid or osmotic pressure by changing fatty acid composition of cell wall, which is termed as cross-protection (Kim, Perl et al. 2001, Meena, Mehla et al. 2016). Based on this fact, we expected that the *L. acidophilus* EG008 strain would exhibit the same phenomenon as well as a significant improvement in heat resistance. However, contrary to expectations, no cross-protection effects were observed, and only thermal resistance was found to be improved (Figure 2C and 2D). This suggests that the increased heat resistance may be due to

reasons other than changes in fatty acid composition. Another possibility is that the three biologically replicated samples were not quantitatively sufficient in their sample size to confirm such difference in cross-resistance, which can be a limitation of the study. Further research will be needed to elucidate the mechanisms for the stress adaptation in the strain.

The second goal of this study was to maintain the beneficial functions of the existing *L. acidophilus* wild-type as much as possible while increasing heat resistance. To investigate this, we constructed complete genome sequences of the developed (EG008) and the wild-type (EG004) strains using the 3rd generation sequencing technology and performed comparative genome analysis. As a result, only two SNPs were found (Figure 4A and 4B) between the sequences (Figure 3A and 3B). In addition, there was no difference between the two strains in the gene annotation and functional analysis, suggesting that the characteristics of the *L. acidophilus* EG004 are maintained. We found three regions encoding Bacteriocin, Acidocin, Enterolysin A, and Helveticin J, in both genomes, which is important feature for probiotic efficacy. Bacteriocin is a proteinaceous or peptidic toxin secreted by bacteria with antibacterial activity against other strains except for itself (Riley 1998, Gálvez, Abriouel et al. 2007). It is well known for its function in the gastrointestinal tract to inhibit the invasion of pathogens or competitors and affect the host's immune system (Dobson, Cotter et al. 2012, Hegarty, Guinane et al. 2016). As these substances are

widely used as biological preservatives due to their high stability in the animals including human, we believe the ability to produce bacteriocin is considered a beneficial property for industrial use of probiotics.

We found two SNPs in the genome of EG008 strain, one of which was a non-synonymous SNP located in murD gene encoding UDP-N-acetylmuramoyl-L-alanine--D-glutamate ligase (Figure 4A). This enzyme is involved in the synthesis of peptidoglycan, a component that strengthens the bacterial cell wall (Bertrand, Auger et al. 1997). When the synthesis of this enzyme is not performed normally or the integrity of the enzyme is destroyed, the cell is dissolved by the turgor pressure inside the cell. Previous studies have shown that the expression of this protein increased when heat stimulation was applied in various strains such as *Staphylococcus aureus* and *Streptococcus thermophilus* (Mengin-Lecreulx, Falla et al. 1999, Li, Bi et al. 2011). This suggests that the expression of UDP-N-acetylmuramoyl-L-alanine--D-glutamate ligase could possibly be the defense mechanism of bacteria against external heat stress. The non-synonymous SNP found in our results changes the 435^{th} amino acid residue of this enzyme from serine to threonine, causing hydrogen loss and acyl gain in side chain (Figure 4A and 4C). We suspected that this change caused a number of changes, such as the volume of the molecule and the location of the hydroxyl groups, affecting the three-dimensional structure and hydrophilicity, thereby altering the resistance to thermal stimulation. In

addition, we confirmed that the genotype of these SNPs was specifically found only in *L. acidophilus* EG008 strain (Figure 4A and 4B). This is one of the evidences that these two SNPs, artificially evolved by the ALE method, can be associated with the improved thermal adaptability.

Another SNP was found in the non-coding region, and there were two genes nearby this SNP (Figure 4B). One of them was a gene encoding Galactose-1-phosphate uridylyltransferase, which is 2 bp away from the SNP located between the core promoter and the ORF starting point. The other was a gene that synthesizes L, D-transpeptidase. This peptidase uses peptidoglycan or its precursor as a substrate to form 3-3 peptidoglycan crosslinks in Gram-positive bacteria (Gupta, Lavollay et al. 2010, Peltier, Courtin et al. 2011). Therefore, there may be mechanisms to regulate the rigidity of the bacterial cell wall by regulating the level of expression of this gene (Brammer, Ghosh et al. 2015). It is also known that L, D-transpeptidase enhances the synthesis of (p)ppGpp alarmone (Hugonnet, Mengin-Lecreulx et al. 2016). The ppGpp induces a stringent response that inhibits RNA synthesis in emergency situations such as heat stress and amino acid shortage (Jain, Kumar et al. 2006). It is known to be a gene involved in the Cpx stress response, one of the well-known envelope damage systems in *E. coli*, and is up-regulated with YgaU when subjected to external stress such as high temperature or high osmotic pressure (Bernal-Cabas, Ayala et al. 2015, Ultee, Ramijan et al. 2019). Based on these

evidences, although it is an intergenic variant, there may be a mechanism that affects the expression level of adjacent genes and ultimately contributes to imparting heat resistance.

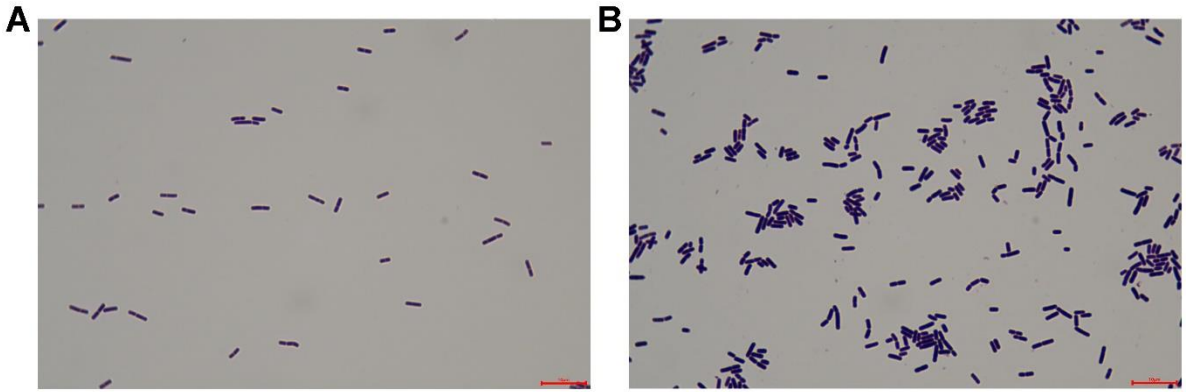
All of the SNPs found in this study were related to cell wall synthesis, specifically the peptidoglycan layer. The cell wall maintains the bacterial cell structure, is responsible for the movement of substance, and interacts with host cells or pathogens (Sequeira, Gaard et al. 1977, Sengupta, Altermann et al. 2013). Thus, since the cell wall is directly affected by various environmental stresses, including thermal stress, the evolutionary response to external stimuli may take precedence. Evidence for this speculation can be found in previous studies investigating changes in cell walls when external physical stresses such as pH, temperature, osmotic pressure, and high bile salt concentration are applied in *L. acidophilus* (Khaleghi, Kermanshahi et al. 2010, Grosu-Tudor, Brown et al. 2016, Palomino, Waehner et al. 2016). Putting all of this together, it indicates that EG008 strain may have evolved to survive at higher temperatures by making cell walls more rigid.

This study has some practical limitations. The first experimental limitation is that the heat-adapted strain selected after ALE treatment may not be the population of strains with the best heat resistance. This occurs by physical restriction in the screening process to isolate single colonies from the population generated through ALE. There are about 10^9 CFUs randomly

mutated strains in the population generated after ALE treatment, but it was not possible to separate the number of all cases into single colonies due to the loss in the dilution process and experimental limitations on culturable colonies on the plate. As a second limitation, although on the whole genome sequences were generated using PacBio long-read sequencing technology, the technical limitation for genotyping error still remained. However, the whole genome we completed had depth coverage of 341.59X for the EG008 strain and 255.60X for the EG004 strain, the possibility of genotyping errors was slim. Thirdly, gene expression analysis was not considered for the comparison of wild-type and heat-adapted strains. Since the genetic variation found in this study can affect the expression level of the transcript, it is expected that this fact will be further revealed if a comparison of the whole transcriptome through RNA sequencing is conducted in the near future. Finally, this study did not cover the experimental validation of the two SNP candidates that were supposed to confer heat resistance to the *L. acidophilus* EG008 strain. Although many gene manipulations using the CRISPR/cas9 system have been reported, many technical difficulties remain in applying CRISPR/cas9 technology to gram-positive bacteria such as LAB (Leenay, Vento et al. 2019). It is our ultimate goal to experimentally verify the mutations detected after ALE and to reveal a direct relationship between phenotypes and genetic factors. We expect that, in the near future, if

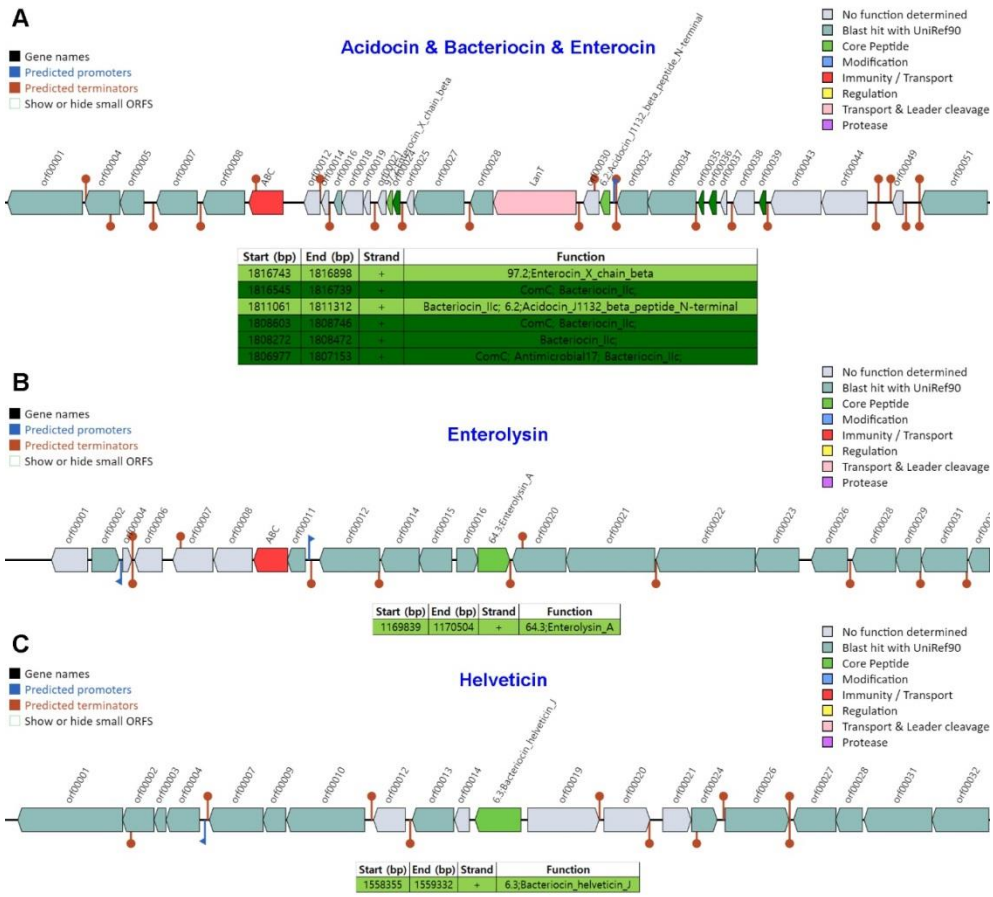
technologies such as genetic scissors for microorganisms become common, the results of this study can be verified.

Despite these technical limitations, we have succeeded in improving our primary target, the heat-resistant limit temperature to 75℃ to maximize the industrial usability of *L. acidophilus* strain. One step further, biomarkers associated with improved thermal resistance was identified through whole genomic analysis. We anticipate the *L. acidophilus* strain developed in this study to be directly helpful in industrial sites where stronger heat resistant bacterial strain is required. We also believe that the identified biomarkers provided insights into the mechanisms of heat resistance and evolution of bacteria, including *L. acidophilus*.



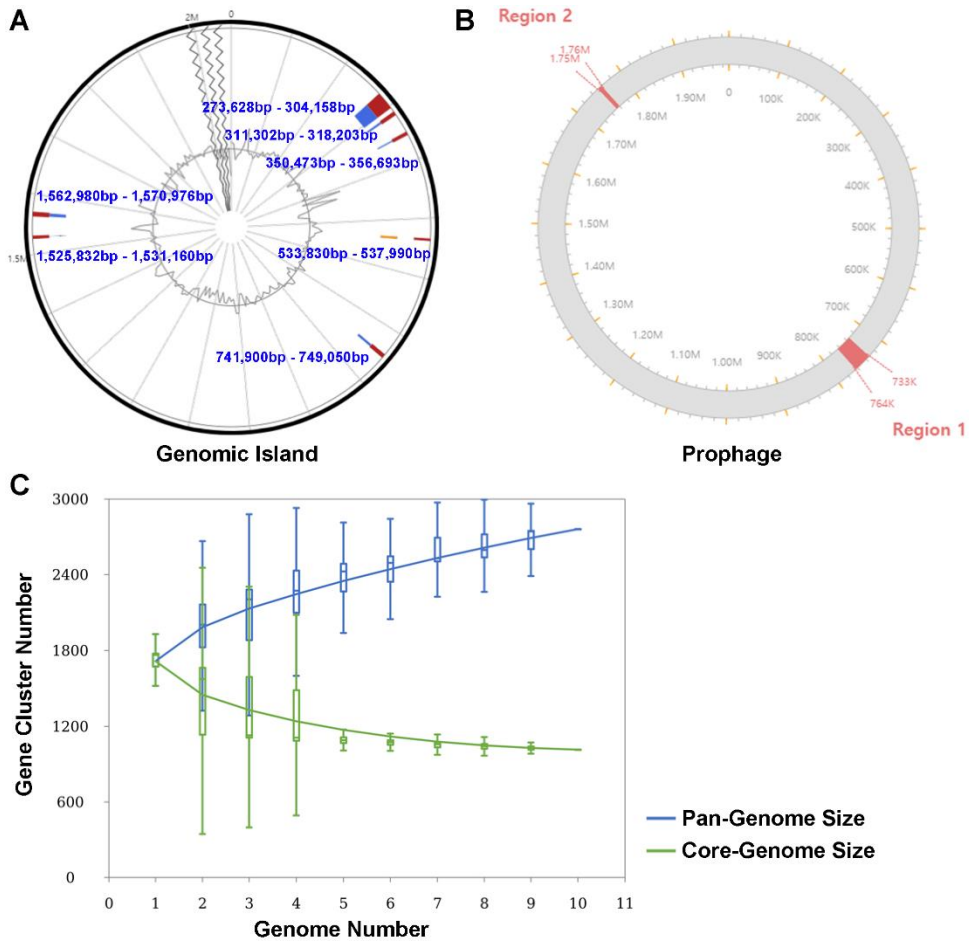
Additional file 10 - Figure S4-1. Staining image of *L. acidophilus* EG004 and EG008 strains.

We observed bacterial morphology of *L. acidophilus* **(A)** EG004 and **(B)** EG008 strains. After 16 hours of incubation, two strains were prepared for microscopic observation. Bacterial morphology was examined under a light microscope at 1000x magnification after gram staining. We verified that both strains are rod-shaped and gram-positive bacteria. The length of the red bar at the bottom represents 10 μm.



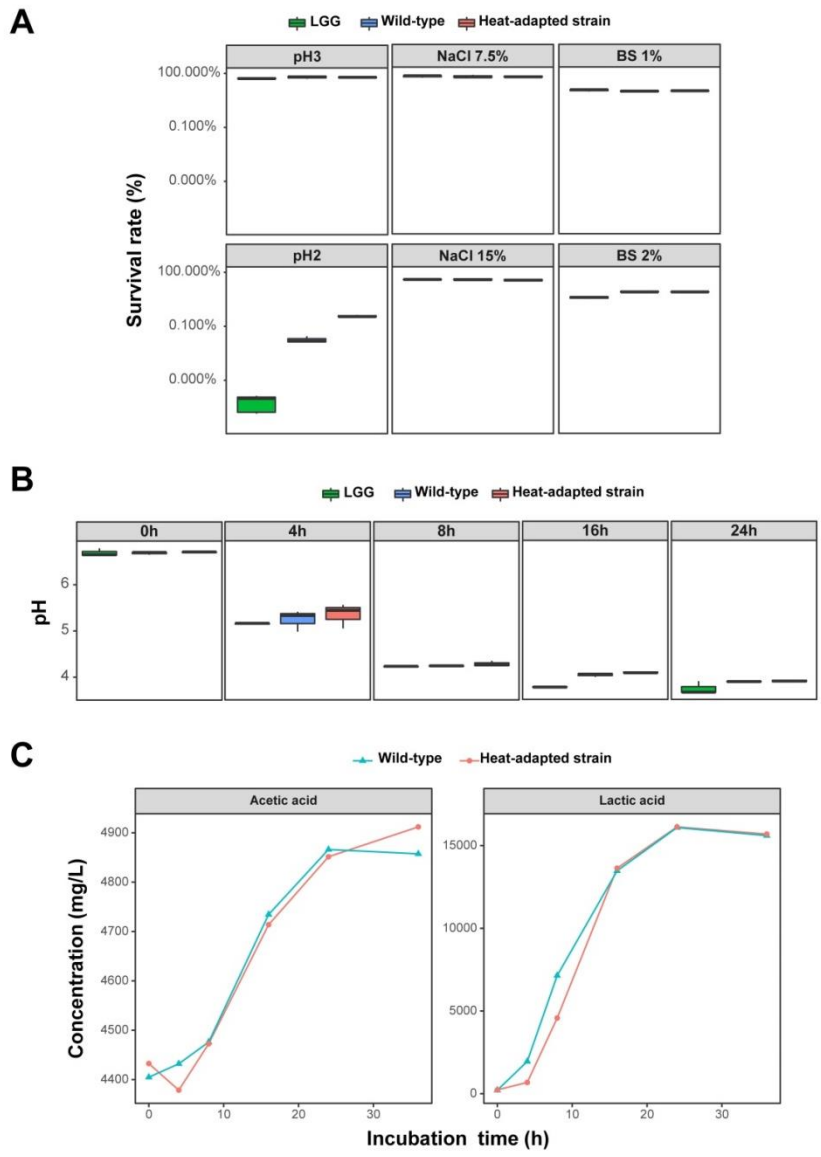
Additional file 11 - Figure S4-5. Antibacterial-related genes found in the genome of *L. acidophilus* EG004 and EG008 strains

We identified that all genes related to antibacterial action, such as Acidocin, Bacteriocin, Enterocin, Enterolysin, and Helveticin, are maintained on the genome of the *L. acidophilus* EG004 using BAGLE4 software. There were no differences between the EG004 and EG008 strains.



Additional file 12 - Figure S4-6. Genomic features of *L. acidophilus* EG004 and EG008 strains.

(A) Seven predicted genomic islands on the *L. acidophilus* EG004 and EG008 strains. **(B)** Two complete prophage regions identified by PHASTER tool. **(C)** Result of pan and core genome analysis. The results demonstrated that the pan-genome of *L. acidophilus* EG004 and EG008 strains is open, given the pattern of influx of new genes while the number of core genes is about half of that of the pan genes. There was no difference between EG004 and EG008 strains in all annotations.



Additional file 13 - Figure S7-4. Physiological activities of *L. acidophilus* EG004 and EG008 strains.

(A) Resistance comparison to external stresses with LGG strain. Resistance was assessed to acid, salt, and bile salt stresses. **(B)** pH change according to bacterial culture time. **(C)** Changes in acetic acid and lactic acid concentration.

Additional file 14 - Table S4-1. Assembly statistic for complete genome of *L. acidophilus* EG004 strain and EG008 strain.

Assembly statistic	***L. acidophilus* EG004**	***L. acidophilus* EG008**
# of subreads	85,438	1,003,160
Avg. alignment length	5,000	3,388
N50 of contig	1,991,561	4,036
Genome coverage	> 100x	> 100x
Maximum contig length (bp)	1,991,559	2,001,616

Additional file 15 - Table S4-2. Microbial kinetics for *L. acidophilus* EG004 and EG008 strains.

Factor	*L. acidophilus* EG004	*L. acidophilus* EG008
μ_{max} (-h)	0.0238	0.0237
$Y_{X/S}$ (cell/substrate)	0.0911	0.0825
t_d (h)	29.08	29.31
K_s (g/L)	2.076	2.079
q (g/h)	0.2616	0.2866

μ: the specific growth rate of the bacteria, μ_{max}: the maximum specific growth rate, S: the concentration of the limiting substrate for bacterial growth, Ks: the half-velocity constant, X: the total biomass, $Y_{X/S}$: fermentation growth yield, t: cultural time, and q: substrate consumption ratio

This chapter was published in *Microbiology Spectrum* (2022)
as a partial fulfillment of Soomin Jeon's Ph.D program.

Chapter 5. Positive Effect of *Lactobacillus acidophilus* EG004 on Cognitive Ability of Healthy Mice by Fecal Microbiome Analysis Using Full-Length 16S-23S rRNA Metagenome Sequencing

5.1. Abstract

The concept of the ‘Gut-brain axis’ has risen. Many types of research demonstrated the mechanism of the GBA and the effect of probiotic intake. Although many studies have been reported, most of the studies are focused on neurodegenerative disease and it is still not clear what type of bacterial strains have positive effects. We designed an experiment to discover a strain that positively affects brain function, which can be recognized through changes in cognitive processes using healthy mice. The experimental group consisted of a control group and three probiotic consumption groups, *Lactobacillus acidophilus*, *Lacticaseibacillus paracasei*, and *Lacticaseibacillus rhamnosus*. All experimental groups fed probiotics showed improved cognitive ability by cognitive-behavioral tests, and the group fed on *L. acidophilus* showed the most. To provide an understanding of the altered microbial composition effect on the brain, we performed full 16S-23S rRNA sequencing using Nanopore, and OTUs were identified at a species level. In the group fed on *L. acidophilus*, the intestinal bacterial ratio of Firmicutes and Proteobacteria phyla increased and the bacterial proportions of 16 species were significantly different from those of the control group. We estimated that the positive results on the cognitive behavioral tests were due to the increased proportion of *L. acidophilus* EG004 strain in the subjects’ intestines since the strain is capable of producing butyrate and therefore modulating neurotransmitters and

neurotrophic factors. We expect that our new strain expands the industrial field of *L. acidophilus* and helps understand the mechanism of the brain-gut axis.

Keywords

Lactobacillus acidophilus, gut microbiome, gut-brain axis, cognitive ability, Nanopore sequencing

5.2. Introduction

The human body is a complex community that habituates various bacteria. Among the bacterial communities in the human body, the gastrointestinal tract is the best bacterial community that has the most abundant and various bacteria (Sender, Fuchs et al. 2016). In 2006, having been released research that obesity is associated with bacterial composition in the gut, a study for gut microbiome began in earnest (Turnbaugh, Ley et al. 2006). The gut microbiome is defined as the collective genomes of microorganisms that live in the gastrointestinal tract. Functions of the gut microbiome have been reported such as nutrient metabolism and regulation of the immune system for the host (Zhang, Tang et al. 2019). Microbial composition in the gut is altered by environmental factors like age, diet, stress, and lifestyle, and the change in microbial composition can induce physical changes in the host (Ghaisas, Maher et al. 2016). In recent, the gut microbiome's effects on the brain have been proved and the concept of the brain-gut axis has risen to the surface (Mayer, Savidge et al. 2014). The brain-gut axis is a complex system involving the enteric nervous system and central nervous system including the brain and spinal cord, and it works with bidirectional communication between the central and the enteric nervous system (Martin, Osadchiy et al. 2018). Although the brain is located apart from the gut, the gut microbiome can affect the brain by stimulating the enteric nervous system and vagus nerve.

Thus, dysbiosis of the gut microbiome often causes brain diseases. The recent experimental results described that gut microbiome dysbiosis was observed in patients with Autism, Alzheimer's disease, and Parkinson's disease (Hill-Burns, Debelius et al. 2017, Pulikkan, Maji et al. 2018, Danilenko, Stavrovskaya et al. 2020, Singhrao and Harding 2020). At the same time, studies on the mechanisms to understand the brain-gut axis have been conducted. First, it was suggested that the microbial-derived metabolites are the main components acting on the neural pathways of the brain-gut axis (Fetissov, Averina et al. 2019, Martin-Gallausiaux, Marinelli et al. 2021). The most well-studied substances are short-chain fatty acids (SCFA) such as acetate, propionate, and butyrate, which are produced in the process of decomposing non-digestible fibers and carbohydrates (Schwiertz, Taras et al. 2010). It promotes indirect signaling to the brain by modulation and induction of neurotransmitter and neurotrophic factors like γ-aminobutyric acid (GABA) and Brain-derived neurotrophic factor (BNDF). Second, the suggestion was that the gut microbiome affects brain function by regulating metabolic pathways (Kaur, Bose et al. 2019). Previous research reported that the level of tryptophan metabolites including serotonin and indolepyruvate was altered by the gut microbiome. These metabolites have roles in the functioning of the gut-brain axis such as signaling and anti-oxidant. Third, the gut microbiome may affect the brain by immune pathway (Miettinen, Vuopio-Varkila et al. 1996). Interferon

(IFN), Tumor necrosis factor (TNF), and Interleukin are well-known immune factors. According to recent studies, the amount of the immune factors is regulated by the intestinal microflora. These immune factors affect brain function by stimulating and activating the hypothalamic-pituitary-adrenal axis. Finally, it was suggested that gut microbes directly influence the brain by altering the fatty acid composition of the brain (Wall, Marques et al. 2012). Several studies have been reported on the correlation between intestinal microbes and the brain, but the specific mechanism of the brain-gut axis is still not clear.

Probiotics are defined as bacteria that have positive effects on the host body (Sanders 2008). Probiotics have been widely used as a health supplement since it has various beneficial functions to host's health with high adhesion property to the intestine and low side effect. Most probiotics include bacteria genera that are gram-positive, facultative anaerobic and rod-shaped. *Lacticaseibacillus rhamnosus* (*Lcb. rhamnosus*) is one of the longest-studied probiotic species, and many strains such as LGG and GR-1 belonging to this genus are commercially available. It is well known that *Lcb. rhamnosus* has healing effects on diarrhea, acute gastroenteritis, and atopic dermatitis (Ö sterlund, Ruotsalainen et al. 2007, Hoang, Shaw et al. 2010, Szajewska, Guarino et al. 2014). Recently, its neurobehavioral effects such as anxiety and depression relief have been reported (McVey Neufeld, O'Mahony et al. 2019). *Lacticaseibacillus paracasei* (*Lcb. paracasei*) is one

of the representative probiotic species, and it has been studied to be effective in treating ulcerative colitis and allergic rhinitis (Ghouri, Richards et al. 2014, Güvenç, Muluk et al. 2016). In a recent study, an effect on age-related cognitive decline and a stress relief effect was reported with several strains of this species (Corpuz, Ichikawa et al. 2018). *Lactobacillus acidophilus* (*L. acidophilus*) is another representative probiotic strain. This strain lowers cholesterol levels and has beneficial health effects such as antibacterial effects against harmful bacteria like *Streptococcus mutans* and *Salmonella typhimurium* (Coconnier, Liévin et al. 1997, Tahmourespour, Salehi et al. 2011).

In this study, we aimed to present a new strain that has an enhancing effect on cognitive ability through the brain-gut axis and provide an additional understanding of the brain-gut axis. Three probiotic strains, *L. acidophilus*, *Lcb. paracasei*, and *Lcb. rhamnosus,* which have previously demonstrated beneficial effects to the host as one of the gut-microbiome strains, were used to confirm their positive effects on cognitive ability. Full 16S and 23S rRNA sequencing was performed to annotate the gut microbiome at a species level for downstream analysis. We expect our results to provide an understanding of the role of the gut microbiome.

5.3. Materials and Methods

5.3.1. Animals

4-week-old C57BL/6 mice (n = 48, average weight 26g) were gained from YoungBio (Seongnam, Korea). Since a male mouse is mainly used in animal experiments for the brain-gut axis and it is estimated that there is no difference between the intestinal environment and brain-gut axis system between female and male, male mice were used for the experiment with reducing the experimental variation. All mice were housed in a group of four per cage under standard controlled laboratory conditions (temperature of 20±5℃, humidity of 55~60%) on a 12-h light/dark cycle (light on at 7:00 a.m.). Each group was constituted of 12 mice, and it was nurtured by distributing 4 mice to 3 cages. Twelve cages were located at random. All animals received *ad libitum* access to food. All animal experiments were performed following protocols approved by the Institutional Animal Care and Use Committee (IACUC) of Seoul National University, and the permission number is SNU-190607-4-3.

5.3.2. Bacterial treatment

The bacterial strains were isolated from fermented dairy foods. When identifying the brain-gut axis effect, the important factors to be considered were viability and adherence capacity. Therefore, we selected the species that are known to have adherence capacity in the GI tract, as well as the

potential for gut-brain axis effect. To identify species of each strain, 16S rRNA genes were sequenced by Macrogen Inc. (Seoul, Korea) with 27F and 1492R primers. Obtained sequences were compared with sequences in the NCBI database using BLAST. The experiment was constituted with 4 groups; 3 experimental groups were fed on autoclaved tap water mixed with *L. acidophilus* EG004, *Lcb. paracasei* EG005, and *Lcb. rhamnosus* EG006, and a control group was fed on sterilized tap water. Each group consisted of 12 mice. Bacteria to delivery were freshly cultivated every day. Probiotic colonies were sub-cultured into 5ml MRS broth for 8 hours. After the sub-culture, 3 probiotic strains were inoculated in 500 ml MRS broth for 16 hours. Cultivated cells were spun down by centrifugation at 4,000 rpm for 10 min. The supernatant was removed, and the pellet was suspended by 0.85 % NaCl solution. Re-suspended cells were centrifuged at 4,000 rpm for 10 min to remove medium ingredients. The washing process was conducted twice. Washed cells were dissolved into autoclaved tap water. The final cell concentration of vehicles was about 1.0E9 CFU/ml. To estimate the probiotics amount per day per subject, daily water intake and probiotic concentration in vehicles were recorded. Cell viability of probiotics was measured by serial dilution and spreading in MRS agar plate. The probiotics amount per day per subject was calculated as an average of daily water intake per subject, by multiplying the average of daily probiotic concentration.

5.3.3. Animal treatment

The animal experiment was designed to minimize animal stress. All animal treatment was described in Figure 1 by timeline. Four weeks old mice were allowed to habituate freely for acclimatization for 1 week. After a week, tap water and water mixed with probiotics were delivered every day. Water intake was monitored every day and body weight was measured every week. Evaluations of cognitive ability were conducted after 4 weeks after probiotic intake. Behavioral tests were conducted at least 2 days after the weight-measurement day to minimize the stress effect. Animals were carried to a behavioral test room to assimilate room conditions and were allowed to relax for 6 hours before any behavioral test. In order to reduce the variance of feeding time, the experimental order of the mice was distributed evenly. All apparatus and objects for the behavioral tests were cleaned with 70 % ethanol and dried after every trial to remove odors and any clues. The mice were sacrificed at the end of 13 weeks after the evaluations of the cognitive behavior. Preliminary experiments were conducted to obtain appropriate experimental values under our experimental environmental conditions. The three to five experimental conditions referring to published results were tested in our laboratory, and the experimental conditions showing a value similar to the average value of the previous studies were determined.

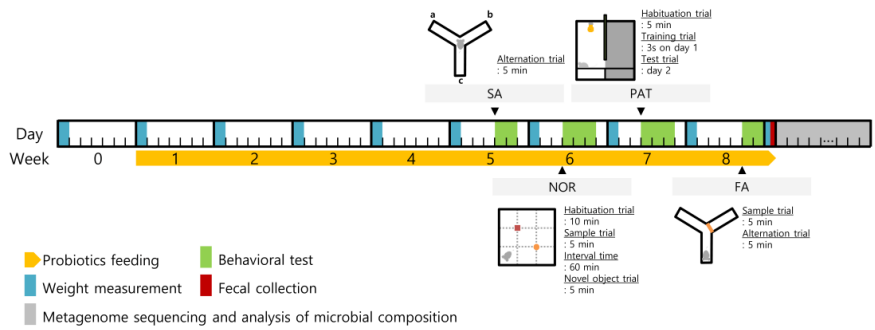


Figure 5-1. Schematic diagram of the study to discover a new probiotic strain with improved cognitive ability

The diagram displays the experimental schedule by day and week for identifying probiotic strain with improved cognitive ability. Cognitive ability was measured once a week by four behavioral tests. The diagram of each experiment shows the first position of the animal.

5.3.4. Y maze (Spontaneous alternation; SA)

Short-term spatial memory was assessed with a Y maze apparatus. SA was used to measure rodents' habit to explore a new environment. The Y maze consisted of 3 identical arms that cross each other with 120° (JEUNGDO Bio & Plant Co., Ltd., Korea). Mice are laid in the middle of the Y maze facing a corner, not an arm. Each animal was allowed to freely navigate all three arms for 5 minutes and the animal's entries to any arm were recorded. An arm entry was determined as any instance when the whole body of the mouse entered the arm and navigated at least 70% of the space. The spatial memory was evaluated by spontaneous alternation, the number of arm entries, and the ratio of mice per group that entered spontaneous alternation during the first three entries. Spontaneous alternation was calculated as shown below.

$$\text{Spontaneous alternation [\%]} = \frac{\text{Number of spontaneous alternation}}{\text{Total number of arm entries} - 2} \times 100$$

5.3.5. Novel object recognition test (NOR)

Based on the concept that mice tend to prefer a new object over a familiar one, a novel object recognition test (NOR test) was performed in an open field (40×40×40 cm (W×D×H), JEUNGDO Bio & Plant Co., Ltd., Korea). Two objects for this test were selected showing similar preferences through the preference test. The test consisted of Sample trial (T1; 10 min), Interval time (IT; 60 min), and Novel object trial (T2; 5 min). In T1, 2

identical objects were located at 1/3 and 2/3 diagonal of the open field, respectively. The animal was laid facing the wall with the same distance to two objects, and was allowed to explore objects for 10 min. After exploration, the mouse came back to the cage and had a rest. In T2, objects were positioned at the same position as T1, but one of the objects was changed to a novel object. To measure the time taken to interact with objects, all experiment processes were recorded, and the exploration time was measured by Movavi software with 3 decimal places. It was recognized as significant only when the mouse approached facing the objects within 2.5 cm. Cases that the mouse climbed objects and individuals with exploration time less than 2 seconds were excluded. The results were presented as a discrimination ratio, the number of object touches, and the ratio of mouse that touched the novel object first before it touched the familiar object. The discrimination ratio was defined as the below equation.

Discrimination ratio [%] =

$$\frac{\text{Novel object interaction time}}{\text{Novel object interaction time} + \text{Familar object interaction time}} \times 100$$

5.3.6. Passive avoidance task (PAT)

The passive avoidance task is designed to evaluate inhibitory avoidance memory according to rodent habit that a mouse prefers dark environment naturally. Shuttle box (41×21×30 cm (W×D×H), JEUNGDO Bio & Plant Co., Ltd., Korea) is an apparatus made for the passive

avoidance task and consists of a bright chamber and a dark chamber which are separated by a sliding door. The floor of the chambers is made of stainless-steel grids to flow current. The test was conducted for 2 days; Acquisition (Day 1) and Test (Day 2). On day 1, a subject was put in the bright chamber facing the wall across the closed sliding door. After the mouse explored the bright chamber for 1 minute, and the moment the mouse was away from the door for over 100 mm, facing the wall not the door, the door was opened so that the mouse could freely enter and move around the dark chamber. Latency time was measured until the mouse entered the dark chamber completely. The door was closed when the animal entered the dark compartment wholly including its tail, and 0.25 mA electric shock was provided to the paws by steel grid for 3 seconds. To memorize the situation, the mouse was kept in the dark chamber for 30 seconds after the shock and returned to the home cage for 24 hours. On day 2, the mouse was laid again into the bright chamber. After 1 minute of adaptation, the sliding door was opened when the mouse faced the wall like day 1. Latency time was measured again until the mouse entered the dark chamber. If the animal rather stayed in the bright chamber for more than 300 seconds (which was the cut-off time), the experiment was completed. All experimental processes were recorded and the time was measured by the Movavi program with 3 decimal places.

5.3.7. Y maze (Forced alternation; FA)

Forced alternation was assessed with the same Y maze as described above. This test consisted of 3 phases; Training trial (T1; 5 min), Interval time (IT; 60 min), and Test trial (T2; 5 min). A mouse was placed at a starting arm of Y maze facing the wall. The subject freely explored the maze during T1, while an entry was blocked with white expanded polystyrene. After the learning trial, the mouse was returned to the home cage and rested for 1 hour. In T2, the mouse was again placed into the starting arm without the plate blocking the novel entry, and explored all three arms. All movements of mice were recorded through video. Forced alternation was evaluated by the ratio of time spent in the novel arm compared to the whole experimental time, time is taken to first enter the novel arm, and the percentage of mice per group that entered the novel arm as their first entry. The case that the mouse passed at 2/3 of the arms was admitted as a valid entrance. An individual that showed no navigation of the maze or that had entered the arms less than 5 times was excluded.

5.3.8. Feces collection and cognitive ability evolution

After all cognitive assessments had been completed, 2-3 stool samples were taken from each experimental subject. Sterilized stainless-steel tweezers were used for fecal picking, tweezers were washed with 70% alcohol and dried sufficiently before collecting new samples. The fresh

samples were immediately enclosed into a 1.5ml Eppendorf tube and were put on ice. Then, it was stored at -80 degrees Celsius until used for 16S rRNA sequencing.

In order to determine the group that showed the best increase in cognitive ability, a score was assigned to the cognitive ability evaluation item. The items used for evaluation are spontaneous alternation, group ratio of SA, discrimination ratio, group ratio of NOR, step latency at day 2, forced alternation, and group ratio of FA (Table S2). Scores were given in ascending order of ranking (1-4 points), and the group with the highest total was selected as the group with the highest cognitive ability increase.

5.3.9. Statistics

Data were analyzed by R studio. Ineligible data were cut based on the requirements mentioned above. Data normality was assessed using the Shapiro-Wilks test and homogeneity of variance was assessed using Levene's test. Wilcoxon rank-sum test and independence t-test was used to evaluate statistical significance between experimental groups. P-values were adjusted by the FDR method for multiple testing corrections. Statistical significance was set as P-value under 0.05. All data are expressed as mean ± SEM.

5.3.10. Full 16S-23S rRNA sequencing

To characterize the microbial community associated with measured cognitive assessment, metagenome sequencing of the 16S-23S rRNA gene was carried out by Oxford Nanopore MinION. Metagenome sequencing was performed for the control group and *L. acidophilus* group, which showed a significant difference from the control in the cognitive ability evaluation. Among the 12 stored stool samples of each group, 5 samples with sufficient amount for sequencing were selected. For library construction, gDNA was extracted from fecal samples using AccuPrep® Stool DNA extraction Kit (Bioneer, Daejeon, South Korea). To identify the quality of extracted gDNA, A260/A280 and A260/A230 absorbance were used with 0.7 % agarose gel electrophoresis. After performing quality control, selected samples were used for the library construction. Stool samples were lysed and bacterial cells were disrupted by Zirconia/Silica Beads and proteinase K. The sequencing library was prepared by 16S-26S rRNA PCR amplification with Nanopore Ligation Kit (SQK-LSK109, Nanopore, Oxford, UK) following the manufacturer's instructions. Purification and quality checks were conducted using agencourt AMPure XP cleanup (Beckman Coulter, CA, USA), Quant-iT^{TM} PicoGreen^{TM} dsDNA Assay Kit (Invitrogen, Ireland), and 0.7% agarose gel. The PCR products were diluted and end-repaired using NEBNext FFPE Repair Mix (New England BioLabs, Ipswich, USA). The amplicon was Nick-repaired using NEBNext End repair/dA-tailing Module (New England BioLabs), prior to adapter ligation by NEBNext

Quick Ligation Module (New England BioLabs). The sequencing library was loaded on primed Flongle flow cell according to Nanopore protocol. Sequencing was performed by MinION MK1b. Sequencing data was acquired by MinKNOW software (19.12.5) without live base-calling. The metagenomic sequences are available in the NCBI database under the accession number PRJNA781018.

5.3.11. Metagenome analysis

Raw data were obtained as fast5 files. Base-calling was carried out by Guppy 4.0.11 with 2,000 chunk size and 4 base callers (Wick, Judd et al. 2019). Porechop version 3 was executed for trimming adapter sequences (https://github.com/rrwick/Porechop). To annotate bacterial taxonomy, trimmed sequences were aligned with MIrROR (http://mirror.egnome.co.kr/) using Minimap2 (Li, Tai et al. 2018). In Operational Taxonomic Unit (OTU) identification, only results with more than 2,500 matching bases and more than 3,500 bases including gaps in mapping were used. To normalize abundance data, the TMM (The trimmed mean of M-values) method was used by the edgeR package of R software (Chen, McCarthy et al. 2014). To characterize each group, biological diversity was calculated through the physeq package of R software (McMurdie and Holmes 2013). A rarefaction curve was constructed to check the saturation of genome sequencing. To compare species richness, alpha diversity was calculated as chao1 and

Shannon indexes. To compare between groups, beta diversity was calculated using Bray-Curtis dissimilarity and Unifrac distance. P-value was calculated by the Adonis test. For detection of unequal features, Wilcoxon rank-sum test was performed in each taxonomic level with 0.95 confidence level. To compare functional profile, PICRUSt2 was performed (Douglas, Maffei et al. 2020). Correlation between cognitive ability and bacterial OTUs was inferred by Spearman's rank correlation analysis. P-values were adjusted by FDR method.

5.3.12. SCFA identification in bacterial culture

To identify the amount of short-chain fatty acids (SCFAs), high-performance liquid chromatography (HPLC) was performed using Ultimate3000 (Thermo Dionex, USA) and Aminex 87H column (300x10mm, Bio-Rad, USA). Bacterial cultures of EG004, EG005, and EG006 were inoculated for 24 hours. After cultivation, the samples were filtered with 0.45 μm of a membrane filter. The filtered sample of 10μL was injected into the HPLC.

5.3.13. Whole-genome sequencing of EG005 and EG006 and Whole-genome sequence of EG004

To identify probiotic safety and potential secondary metabolite producing ability, whole-genome sequencing of *Lcb. paracasei* EG005 and

Lcb. rhamnosus EG006 was performed. For library construction, DNA was extracted from cultured bacterial cells. After performing quality control, gDNA was used for the library construction. Bacterial cells were lysed by lysozyme for gram-positive bacteria, and removed RNA and protein to isolate DNA. Quality control for gDNA was conducted by 260/280, 260/230 absorbance with 0.8% agarose gel. Genomic DNA was fragmented to a target length of 20Kb using g-Tube (Covaris, MA, USA) and Short DNA fragments <5 kb are depleted by SRE (Circulomics, MD, USA). The fragments were End-prepared, Nick-Repaired, and then ligated with Nanopore adapter. After every enzyme reaction, the DNA samples were purified using AMPure XP beads (Beckman Coulter, CA, USA) and QC with Quant-iT™ PicoGreen™ dsDNA Assay Kit. The sequencing library was loaded on primed Flongle flow cell according to Nanopore protocol. Sequencing was performed on a MinION by MinKNOW software.

Base-calling from raw data was conducted by Guppy Basecaller v4.0.15 with filtering with an average basecall Phred quality score. Adapter sequences were trimmed by PoreChop v0.2.4. Genome assembly was conducted by Canu. Assembled contigs were polished by Nanopolish and racon, and pilon. Circlator circularized each contig and detect replication origin. Assembled contig was assessed by BUSCO 3.0.2. The complete sequences of *Lcb. paracasei* EG005 and *Lcb. rhamnosus* EG006 is available in the NCBI database with accession numbers, SAMN23227569 and

SAMN23227570, respectively. The complete sequence of *L. acidophilus* EG004 that is deposited in the NCBI database with accession number PRJNA657145 was used (Jeon, Kim et al. 2021).

5.3.14. Comparative analysis of bacterial genome sequences

Genetic map was generated by CGView server (Grant and Stothard 2008). To check safety and functionality as probiotics, genetic factors were identified by whole-genome sequences. Virulence factor and prophage gene were detected by VirulenceFinder 2.0 and PHASTER, respectively. IslandViewer4 identified genomic island and crisprfinder searched CRISPR region. Bacteriocin detection was conducted by BAGLE4. To compare functional gene contents, protein prediction was performed by the RAST server. Predicted protein sequences were classified by the SEED system. Categorized protein sequences showed as the proportion in the total predicted sequences.

5.3.15. Data availability

The complete sequences of *Lcb. paracasei* EG005 and *Lcb. rhamnosus* EG006 are available in the NCBI database with accession numbers, SAMN23227569 and SAMN23227570, respectively. The metagenomic sequences are available in the NCBI database under the accession number PRJNA781018.

5.4. Results

5.4.1. Bacterial and animal treatments

Three probiotic strains, *L. acidophilus* EG004, *Lcb. paracasei* EG005, and *Lcb. rhamnosus* EG006, have been identified by the 16S rRNA sequencing. These strains were clustered with available *L. acidophilus*, *Lcb. paracasei*, and *Lcb. rhamnosus* strains, respectively, in a phylogenetic tree of 16S rRNA gene (Figure S1). Probiotic strains were consumed by mice for 8 weeks with assessments of cognitive ability (Figure 1). The averages of daily water intake per subject were similar between groups (Figure 2A). Daily probiotic intakes were maintained constantly and the average amount of *L. acidophilus* group, *Lcb. paracasei* group, and *Lcb. rhamnosus* group were calculated as (7.82E09 ± 1.95E09), (4.37E10±5.17E09), and (3.74E10±3.98E09) CFUs (Figure 2B). To identify the additional effect of probiotics, the body weights of mice were measured every week (Figure 2C and S2). Patterns of weight gain in the 4 groups were similar for 8 weeks. The mean body weight gains of the control group showed the highest value, which was 9.08 g. *Lcb. paracasei* group showed a significant difference from the control group with P-value under 0.05 in the second measurement, but the difference was immediately recovered. Similar to weekly weight change, statistical significance was not found in accumulated weight between experimental groups for 8 weeks.

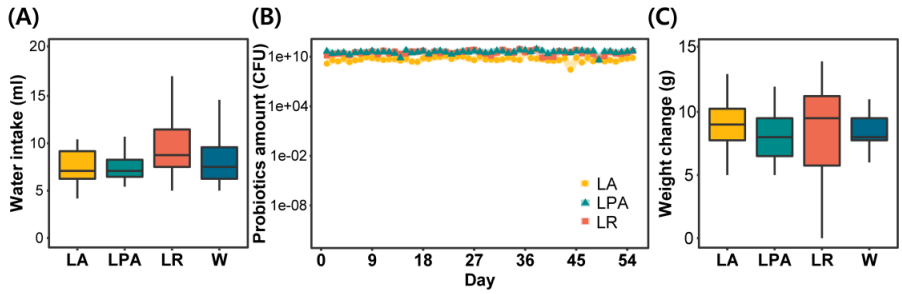


Figure 5-2. Measurement of additional effect after probiotic consumption

Experimental groups are expressed in abbreviations. LA: *L. acidophilus* group, LPA: *Lcb. Paracasei* group, LR: *Lcb. Rhamnosus group*, and W: tap water-fed group (control). **(A)** The average daily water intake. All groups showed a similar average. **(B)** The change of daily intaken probiotic amount by timeline. *L. acidophilus* was ingested in smaller amounts compared to the other two strains. **(C)** The average body weight change for 8 weeks. All groups showed similar averages.

5.4.2. Cognitive behavioral tests

Spontaneous alternation test was conducted to assess spatial learning and short-term memory. Although the average number of the total entries to each arm in *Lcb. paracasei* group was slightly low, the difference between groups was not found (Figure 3A). The comparison of the mouse ratio showed spontaneous alternation for the first 3 entries, *L. acidophilus* group showed the highest value as 75.0%. (Table S1). In spontaneous alternation, the average values of probiotics-fed groups were higher than the value of the vehicle-fed group (Figure 3B). Among the 4 experimental groups, *L. acidophilus* group showed the highest alternation ratio. Wilcoxon rank-sum test was performed to identify statistical significance, but there was no statistical difference between the experimental groups and control group.

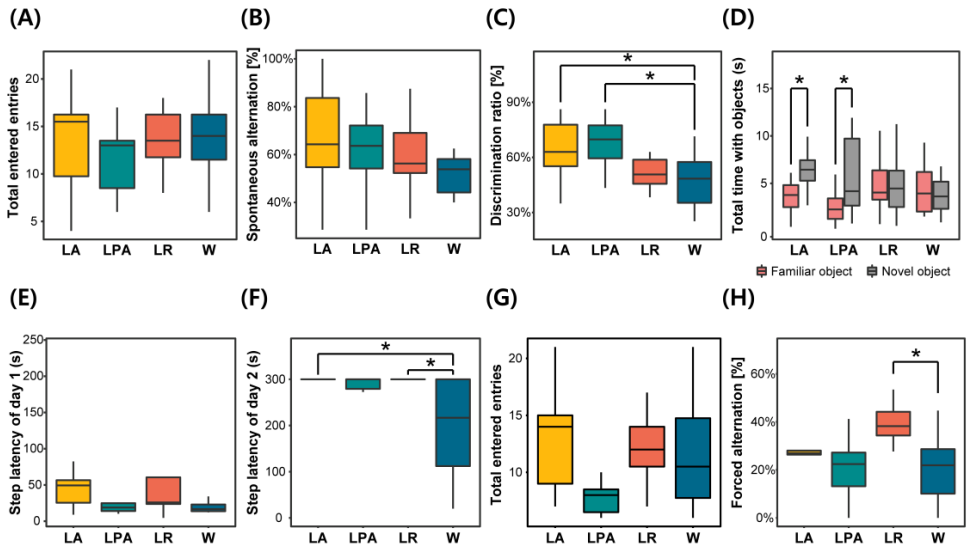


Figure 5-3. Results of cognitive behavioral tests

Experimental groups are expressed in abbreviations. LA: *L. acidophilus* group, LPA: *Lcb. Paracasei* group, LR: *Lcb. rhamnosus* group, and W: the group fed on tap water (control). (**A)** Total arm entries during spontaneous alternation test. **(B)** Spontaneous alternation. This is the representative value of spontaneous alternation test. **(C)** Discrimination ratio. It is the representative value of the novel object recognition test. **(D)** Comparison of the total time to observe two objects. **(E)** Step-through latency of day 1. **(F)** Step-through latency of day 2. This is the representative result of the passive avoidance task. **(G)** Total arm entries during forced alternation test. **(H)** Forced alternation. This result is a representative value of forced alternation. All comparison of average between experimental groups was measured by Wilcoxon rank-sum test. Significant difference is presented with symbol (Adjusted P-value* < 0.05).

Novel object recognition (NOR) test was performed to evaluate long-term and explicit memory using 4 different features (Figure 3C, 3D, and Table S1). *L. acidophilus* group exhibited the highest average ratio of mouse that touched the novel object before the familiar object, whereas *Lcb. rhamnosus* group showed the lowest value under the control group. At discrimination ratio comparison, the three probiotics-fed groups showed higher average values than the control, and *L. acidophilus* group showed the highest values. To identify if there is a significant difference, Wilcoxon rank-sum test was performed. When compared to the vehicle-fed group, *L. acidophilus* and *Lcb. paracasei* groups displayed statistically significant differences with the adjusted P-value of 0.037. To identify animal behavior detail, the number of objects touch and the total time of object observation in each group were compared. In a comparison of object touch, statistical differences were significant in *L. acidophilus* and *Lcb. paracasei* groups with P-values of 0.031 and 0.042, respectively. Also, *L. acidophilus* group had a significant difference between the time taken to observe the familiar object and the novel object.

Passive avoidance task was conducted to measure long-term and implicit memory. Step-through latency was used to compare the mean difference between the experimental groups. Most of the subjects were transferred into a darkroom for a minute on day 1 (Figure 3E). Only 3 animals took over 100 seconds to get into the darkroom. The difference

between the experimental group and the control was not found on day 1. When compared to the latency time on day 1, the average latency time increased on day 2, and unexpectedly, 26 animals stayed in the lightroom for over 300 seconds (Figure 3F). *Lcb. rhamnosus* group presented the highest average latency time, followed by *L. acidophilus* group while the control group showed the lowest average (Table S1). The Mann-Whitney U test was conducted to check the mean difference, the P-values of *L. acidophilus* and *Lcb. rhamnosus* groups were less than 0.05 compared to the control group. The adjusted P values of both groups were 0.040.

To assess spatial learning and long-term memory, forced alternation was conducted. Memory was evaluated by forced alternation (%), the number of arms that the mouse entered, and the percentage of mice in a group that entered the novel arm as their first entry. While the total number of the entries into each arm was diverse, there was no significant difference between the experimental groups and control (Figure 3G). *L. acidophilus* group scored the highest ratio of mice entered the novel arm as their first entry (Table S1). Forced alternation values of *L. acidophilus* and *Lcb. rhamnosus* groups were higher than the value of the control group (Figure 3H). Forced alternation of *Lcb. rhamnosus* group and the control group had a significant difference with the adjusted P-value of 0.038.

Table 5-1. Metagenomic sequencing statistics of *L. acidophilus group* and control

	The number of samples	**Total number of reads**	**Estimated base (Mb)**	**N50**	**Total number of counts**	**Total number of OTUs**
LA[a]	5	312,384±31,887	1,434±143	4,872±90	252401.6±25,171	528.4±40
W[b]	5	335,356±45,814	1,485.6±215	4,748±40	259945.6±35,117	539.8±25
Total	10	323,870±37,604	1,459.8±173	4810±72	256173.6±28,860	534.1±32

[a]: *L. acidophilus* group, [b]: control group. There was no significant difference between groups. All values were presented as average ± standard error of the mean. Fecal samples compiled after 8 weeks of probiotic ingestion were used for metagenome sequencing.

5.4.3. Full 16S-23S rRNA sequencing and biological diversity

Metagenome sequencing was performed with *L. acidophilus* and control groups, which showed the most improvement in cognitive ability. We compared the microbial composition of both groups. Gut microbial component information annotated at a species level was completely constructed by sequencing the entire 16S-23S rRNA of the mouse stool (Table 1). Averagely, 323870.0±84085.5 reads were generated from 10 stool samples. The total number of identified OTU was 252401.6±56284.7 in *L. acidophilus* group and 259945.6±78526.0 in the control group. The produced OTUs were annotated as a total of 528.4±90.4 species in *L. acidophilus* group and 539.8±55.4 species in the control group. To check the sufficiency of the sequencing depth for the analysis, a rarefaction curve was created (Figure 4A).

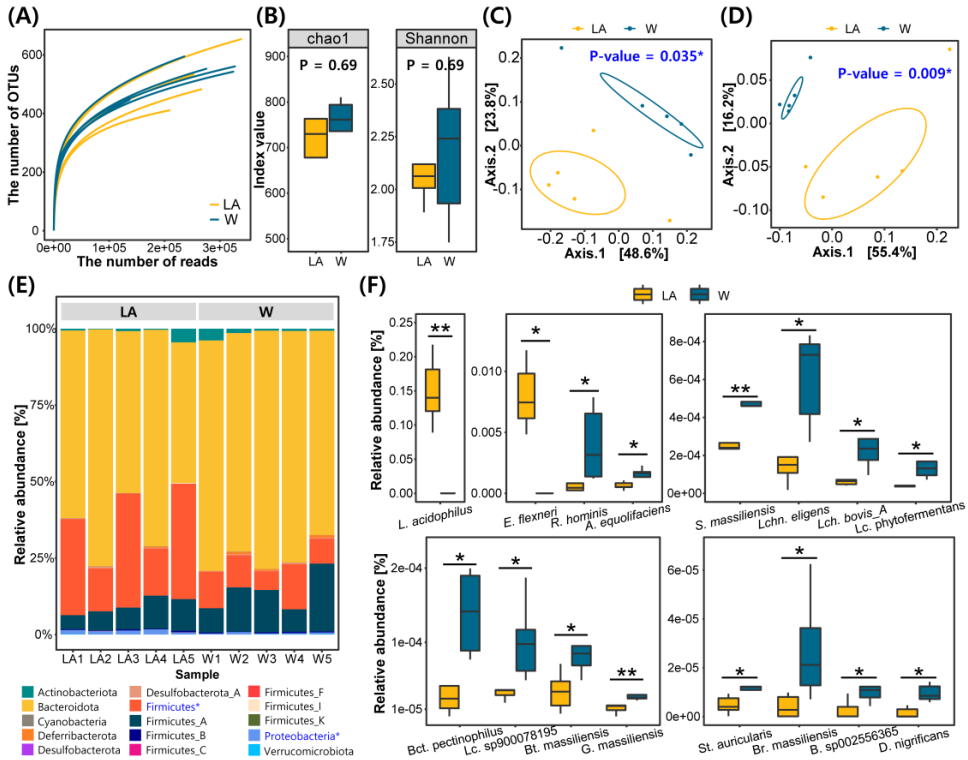


Figure 5-4. Results of metagenomics sequencing

Experimental groups are expressed in abbreviations. LA: *L. acidophilus group* and W: the group fed on tap water (control). **(A)** Rarefaction curve of metagenome sequencing. **(B)** Alpha-diversity of the *L. acidophilus* group and control. **(C)** Beta-diversity using Bray-Cutis distance between the *L. acidophilus* group and control. **(D)** Beta-diversity using Unifrac distance between both groups. **(E)** Comparison of microbial composition at the phylum level. The blue-colored phylum with the (*) symbol showed a significant difference compared to the two experimental groups. **(F)** Comparison of microbial composition at the species level. L. acidophilus: *Lactobacillus acidophilus*, E. flexneri: *Escherichia flexneri*, R. hominis: *Roseburia hominis*, A. equolifaciens: *Adlercreutzia equolifaciens*, S. massiliensis: *Soleaferrea massiliensis*, Lchn. Eligens: *Lachnospira eligens*, Lch. Bovis_A: *Lachnobacterium bovis_A*, Lc. Phytofermentans: *Lachnoclostridium phytofermentans*, Bct. Pectinophilus: *Bacteroides_F pectinophilus*, Lc. Sp900078195: *Lachnoclostridium*

sp900078195, Bt. Massiliensis: *Bittarella massiliensis*, G. massiliensis: *Gemella massiliensis*, St. auricularis: *Staphylococcus auricularis*, Br. Massiliensis: *Bariatricus massiliensis*, B. sp002556365: *Bacillus_AW sp002556365*, D. nigrificans: *Desulfotomaculum nigrificans.* All comparisons of average between experimental groups were measured by independence t-test. Significant difference is presented with symbol (Adjusted P-value* < 0.05, P-value** < 0.01).

Alpha diversity was calculated to compare species richness within a group (Figure 4B). In the comparison of the two groups, no significant difference was found in Chao1 Shannon indexes. Beta diversity was measured to compare the diversity of the microbial community between the two groups (Figure 4C and D). It was confirmed that both beta diversity evaluations (Bray-Curtis and Unifrac distance) had significant differences.

5.4.4. Microbial composition

In the comparative analysis of microbial compositions, taxonomies with significantly different ratios were found between *L. acidophilus* group and the control group. At the phylum level, Bacteroidota accounted for the highest proportion in both groups, followed by Firmicutes (Figure 4E). Significant differences between the two groups were found in 2 of the 12 phyla (Firmicutes, Proteobacteria), all of which were high in *L. acidophilus* group. At the class level, Bacteroidia showed the highest proportion in both groups. Also, the proportion of Bacilli and Gammaproteobacteria classes were increased in *L. acidophilus* group when compared to the control group (Figure S3). At the order level, Bacteroidales showed the highest percentage in both groups, and Lactobacillales and Enterobacterales orders were found to exhibit higher proportions in *L. acidophilus* group. At the family level, *Muribaculaceae* showed the highest proportion in both groups. It was found that 2 families (*Lactobacillaceae* and *Enterobacteriaceae*) showed increased

proportions in *L. acidophilus* group, while a decreased percentage was observed in one family (*Ruminococcaceae*). In the Genus comparison, *Muribaculum* genus showed the highest ratio in the two groups, and 12 genera showed differences between groups. Three genera showed an increased proportion in the experimental group, whereas 9 genera showed higher mean values in the control group. The genus increased in *L. acidophilus* group were *Lactobacillus*, *Staphylococcus_A*, and *Escherichia*, whereas the genera decreased in *L. acidophilus* group were *Bacteroides_F*, *Desulfotomaculum*, *Lachnobacterium*, *Bittarella*, *Agathobacter*, *Roseburia*, *Bariatricus*, and *Lachnospirarea*. At the Species level, *Muribaculum intestinale* was found to account for the largest proportion, with over 50% in both groups. Following *M. intestinale*, the species *Lactobacillus acidophilus*, *Lactobacillus johnsonii*, *Lactobacillus_B murinus*, and *Lactobacillus_H reuteri* were found with a high proportion in *L. acidophilus* group, while *Lactobacillus_B murinus*, *Bacteroides_B vulgatus*, *Faecalibaculum rodentium*, and *Kineothrix alysoides* species showed a high proportion in the control group. No unique bacterial species were found in either of the two groups. Seventeen species showed differences between groups, and it was confirmed that the proportions of *L. acidophilus* and *E. flexneri* were increased in *L. acidophilus* group (Figure 4F).

5.4.5. Functional profiling and correlation analysis

Functional profiling was performed at the KEGG level 3 to estimate the effect of the differential composition of intestinal microbes on the mice (Figure 5). By calculating the LDA score, it was confirmed that the two groups showed significantly different patterns in 9 categories. All nine categories were predicted to be more activated in *L. acidophilus* group. The Phosphotransferase system (PTS) scored the highest, followed by *Staphylococcus aureus* infection, Synthesis and degradation of ketone bodies.

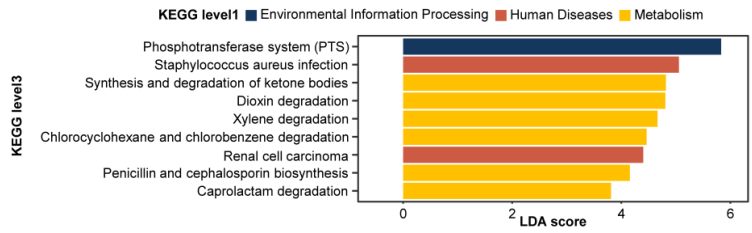


Figure 5-52. Results of functional profiling

Predictive functional profiling of microbiome. All predicted functions have a positive LDA score for the *L. acidophilus* group.

To further estimate the influence of the altered gut microbiota, Spearman's correlation analysis of cognitive-behavioral abilities and bacterial OTUs, and fermentation products were performed (Figure 6). *L. acidophilus* and *E. flexneri* showed a positive correlation with all assessments of cognitive abilities, while the other 14 OTUs presented a negative correlation. In particular, step-through latency at Day 2 and Step latency difference for 2 days of the PAT results showed a significant negative correlation with the *Gemella massiliensis* ($r = -0.8379$, $p = 0.03248$ and $r = -0.8182$, $p = 0.0376$) and *Desulfotomaculum nigrificans* ($r = -0.8781$, $p = 0.01914$ and $r = -0.8450$, $p = 0.03225$).

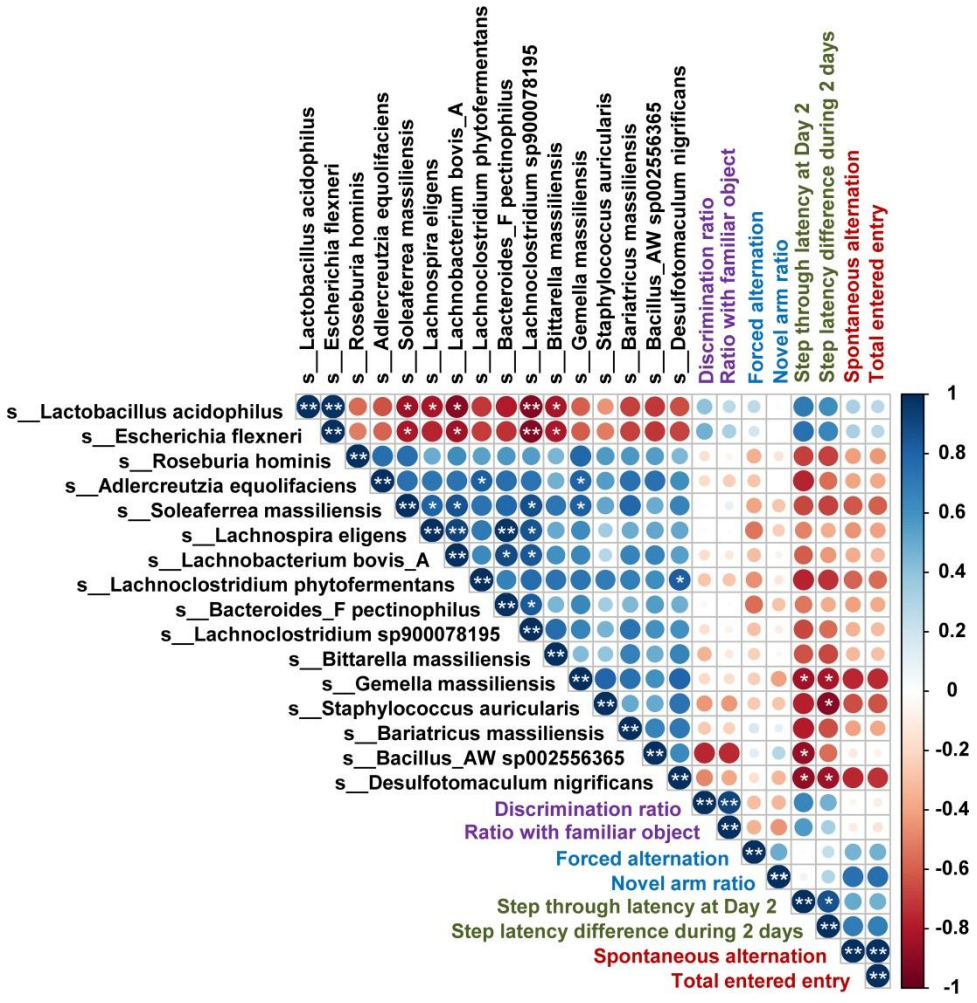


Figure 5-6. Spearman's rank correlation analysis

Correlation analysis was conducted to detect association among bacterial OTUs, measured cognitive abilities, and fermentation products. The color intensity and circle size show the strength of the correlation. Red color represents a negative correlation, and blue color is a positive correlation. Only circles with adjusted P-value under 0.01 are illustrated in the matrix. Results of cognitive ability evaluation were classified by 4 colors: NOR (purple), FA (blue), PAT (deep green), and SA (brown). Significant P values indicated by the symbol * (<0.05) and ** (<0.01).

To provide evidence to indirectly infer the mechanism of action of the gut microbiome, the concentration of SCFA in the microbial culture was measured (Table S2). Lactic acid and acetic acid were found in three microbial cultures. Lactic acid was identified in the highest concentration in *Lcb. paracasei* EG005, and acetic acid was included in the highest concentration in *L. acidophilus* EG004 culture. Propionate and butyrate were not within detectable ranges.

5.4.6. Comparative analysis of genetic contents in bacterial whole-genome sequences

To identify its safety and functionality, several genetic factors were detected. Fourteen genomic islands, two prophage regions, one CRISPR region, and three bacteriocins were found in the genome of *L. acidophilus* EG004. In *Lcb. paracasei* EG005, 29 genomic islands, 7 prophage regions, 3 CRISPR regions, and 2 bacteriocins were detected (Figure S4-S6). In the case of *Lcb. rhamnosus* EG006, 23 genomic islands, 8 prophage regions, 3 CRISPR regions, and 1 bacteriocin were found in the genome. To estimate a genetic factor related to cognitive ability, protein annotation was conducted (Figure 7A). Protein metabolism, Carbohydrates, Amino acids and derivatives showed high proportions, but there was a difference in order by bacterial strains. Protein metabolism had the highest proportion in *L. acidophilus* EG004 and carbohydrates presented the highest proportion in

Lcb. paracasei EG005 and *Lcb. rhamnosus* EG006. In a subcategory comparison of predicted functional sequences, a difference of genetic contents was found (Figure 7B). CDSs related to Fatty acids were found in the genomes of *Lcb. paracasei* EG005 and *Lcb. rhamnosus* EG006. Genes of 3 subcategories (Aromatic amino acids and derivatives, Alanine, serine, and glycine, and Proline and 4-hydroxyproline) were detected in *Lcb. rhamnosus* EG006, while genes of 3 other categories in Amino Acids in Derivatives were contained in only *L. acidophilus* EG004.

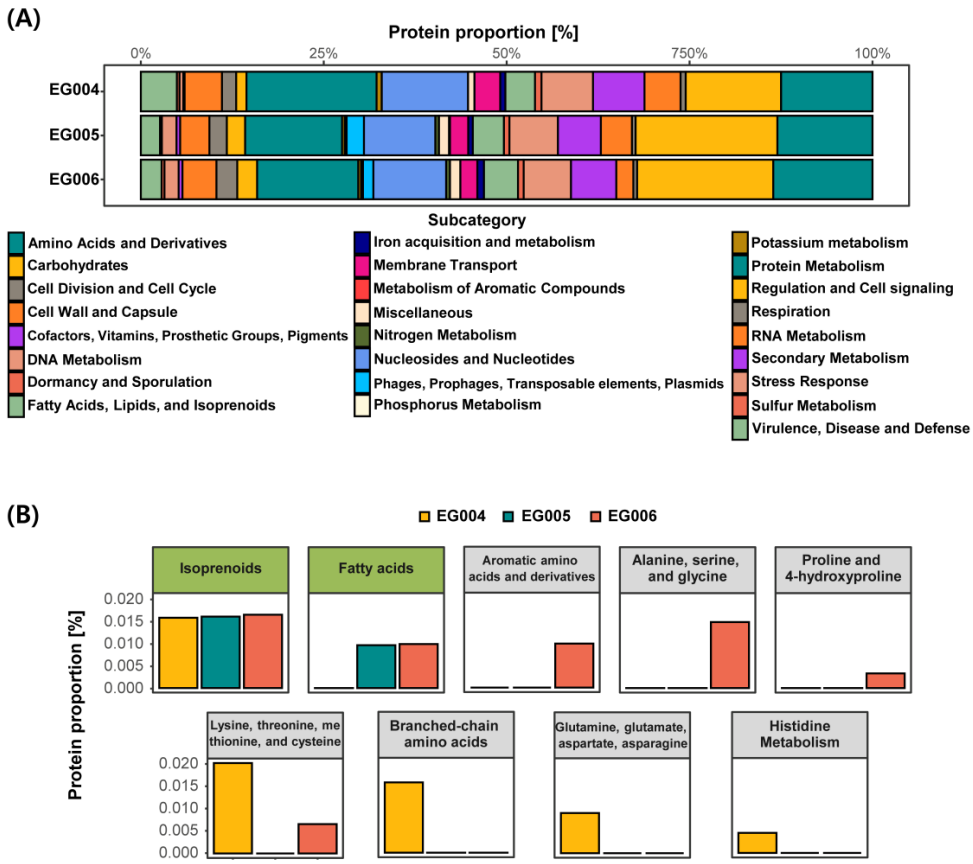


Figure 5-7. Genomic comparison of 3 probiotic strains

(A) Functional classification of protein coding sequences. All predicted protein sequences were classified by categories by SEED system. (B) Subcategories in [Fatty Acids, Lipids, and Isoprenolds] and [Amino Acids and Derivatives]. [Fatty Acids, Lipids, and Isoprenolds] subcategory showed yellow-green colored head and [Amino Acids and Derivatives] category presented light gray colored head.

5.5. Discussion

As interest in Gut-Brain Axis has increased, many types of research in this criterion have been published. However, it is still unclear about the integral mechanism and which strain has a positive or negative effect. Therefore, we aimed to develop a new strain that has a positive effect on the host's cognition, and we found 3 strains that caused positive effects in 4 different cognitive tests (Figure 3). All experimental groups fed on the probiotic strains appeared to improve cognitive ability. The group fed on *L. acidophilus* showed the highest score with a total score of 26, while the groups fed on *Lcb. rhamnosus* and *Lcb. rhamnosus* scored 16 and 15 points respectively, slightly higher than the control group (Table S3). *L. acidophilus* group showed high scores in all evaluations, while others were highly evaluated only in some experiments. In addition, although probiotic consumptions were carried out as the same method, three experimental groups showed improved cognitive ability in different tests. It implies that different probiotic strains affect cognitive ability by different mechanisms, and that *L. acidophilus* had an effect on a wider area than other strains. *Lcb. paracasei* group showed improved cognitive ability in the novel object recognition test. Previous studies indicated that this bacterium improves cognitive ability and increases the level of serotonin and BDNF in the hippocampus (Corpuz, Ichikawa et al. 2018). Other strain, *Lcb. rhamnosus*, displayed improved cognitive ability in passive avoidance task and forced

alternation test. Several studies demonstrated that *Lcb. rhamnosus* consumption could increase cognitive ability by activating microglia in the hippocampus (Huang, Chen et al. 2018, Wang, Ahmadi et al. 2020). Similar to previous studies, we experimentally confirmed that *Lcb. paracasei* and *Lcb. rhamnosus* could enhance cognitive function. On the other hand, although it is indicated that *L. acidophilus* strain has a neuroprotective effect against traumatic brain injury, there was no experimental research related to its cognitive ability (Kelly, Allen et al. 2017, Oh, Joung et al. 2020). In our study, we identified that *L. acidophilus* group presented the highest classical measured values as well as incidental measured values in novel object recognition tests and passive avoidance tasks. This indicates that *L. acidophilus* is capable of improving cognitive ability comparable to that of previously reported strains. Our results will help further broaden the industrial field of probiotic strains.

To understand the effect of the gut microbiome on the brain as our secondary goal, we performed gut microbiome analysis of *L. acidophilus* group, which showed the best cognitive improvement, along with the control group. The difference of species richness was not found in the comparison of alpha diversity, whereas the difference was found in the comparison of beta diversity (Figure 4B, 4C, and 4D). It represents that the number of OTUs constituting the two gut microbial communities is similar, but the composition of the OTUs is different. In the comparison of the two

communities, significant differences were observed at all taxonomic levels except for the bacteria kingdom, which was mostly *L. acidophilus.* Naturally, *L. acidophilus* group was confirmed to show a significant increase in *L. acidophilus* abundance and ultimately show a high ratio of *L. acidophilus.* This indicates that a large amount of *L. acidophilus* is capable of safely reaching the intestines without being affected by digestive juices such as gastric acid and pancreatic enzymes.

We estimated that the positive effect on cognitive ability due to the increased proportion of *L. acidophilus* in the intestines was based on two rationales: modulation of neurotransmitters and neurotrophic factors and production of SCFAs. First, *L. acidophilus* modulates several types of neurotransmitters in the intestine. Microbial-derived intermediates, which affect the brain through gut epithelial and blood-brain barriers, are such as GABA (γ-aminobutyric acid), glutamate, dopamine, noradrenaline, serotonin (5-Hydroxytryptamine; 5-HT), and Brain-derived neurotrophic factor (BNDF). These neurotransmitters are synthesized from various amino acids. GABA and glutamate are produced from the gut microbiome such as *Bifidobacterium* and *Lactobacillus* (Yunes, Poluektova et al. 2016). Glutamate has a role as a neurotransmitter by itself, and it is used at GABA synthesis (Walls, Waagepetersen et al. 2015). Dopamine and Noradrenaline are synthesized from specific amino acids such as tyrosine and phenylalanine (Lehnert, Wurtman et al. 1993). L-Tryptophan is a well-

known precursor of serotonin (O'Mahony, Clarke et al. 2015). Therefore, altered amino acid composition by the gut microbiome seems to affect the host's neurotransmitter synthesis. In the comparison of the functional protein genes, *L. acidophilus* EG004 showed a higher composition of the gene related to amino acid metabolism, than *Lcb. paracasei* EG005 and *Lcb. rhamnosus* EG006 showed (Figure 7A). Changes in intestinal amino acid composition caused by ingested *L. acidophilus* may have led to differences in cognitive ability. It has been proven that *L. acidophilus* consumption produces and up-regulates neurotransmitter and neurotrophic factors including GABA and serotonin (Lim, Kim et al. 2009, Lim, Yoo et al. 2009, Cao, Feng et al. 2018, Rahimlou, Hosseini et al. 2020). Thus, it is estimated that increased *L. acidophilus* EG004 in the gut modulates neurotransmitters and affects the animal's nerve system. Second, SCFAs, fermentation products of *L. acidophilus*, positively apply to brain function. For example, acetate, one of the short-chain fatty acids (SCFAs), promotes the activation of the parasympathetic nervous system (Perry, Peng et al. 2016). Also, it is indicated that acetate improved cognitive ability and neurogenesis in the hippocampus with increasing BDNF and IGF-1 levels as a glatiramer acetate form (He, Zou et al. 2014). Likewise, butyrate, a famous HDAC inhibitor, has been used for pharmacological purposes since lower global histone acetylation is a common phenomenon observed in many neurodegenerative diseases (Bourassa, Alim et al. 2016). Its therapeutic

effect on neurodegenerative diseases including Parkinson's disease was verified, showing enhancement of neurotrophic factors and improvement in learning and memorizing (Barichello, Generoso et al. 2015). However, SCFAs are not produced until non-digestible carbohydrates reach the small intestine to be broken down by microbial metabolism, so it is not fully produced by the human digestive enzymes without specific microbes. *L. acidophilus* is a representative species that produces SCFAs through non-digestive carbohydrates, and it can be assumed that the intake of *L. acidophilus* EG004 caused the increase in SCFAs of the experimental mice's gut. The result of SCFA measurement in bacterial culture raises the possibility of this assumption (Table S2). Although it is different from the metabolism in the gut since the SCFAs were measured in the medium to which glucose is the main energy source, it indirectly estimates its SCFA-producing ability. The result of functional profiling in our study also upholds this (Figure 5B). In the analysis of functional profiling, activation of genes of synthesis and degradation of ketone bodies was predicted by comparing it with control. The ketone body is one of the main fuels of the brain like lactate and butyrate, which is the main product of *L. acidophilus*, and is also capable of replacing glucose as an alternative fuel. Similar to butyrate mentioned earlier, ketone bodies modulate the brain with anti-oxidant reaction, energy supply, regulation of deacetylation activity, and regulation of the immune system. In recent studies, it is indicated that the

increase of ketone body's concentration induces an alleviation effect on brain diseases such as epilepsy, Alzheimer's disease, and Parkinson's disease as well as memory improvement (Klein, Janousek et al. 2010, Hertz, Chen et al. 2015, Norwitz, Hu et al. 2019). Based on this evidence, ingested *L. acidophilus* EG004 in our experimental group seems to have produced SCFAs and modulated neurotransmitters, and *L. acidophilus*-derived metabolite would have raised cognitive ability. Although we did not measure microbial-derived metabolites, previous researches demonstrated that probiotic consumption leads to an increase of microbial-derived metabolites in the intestines.

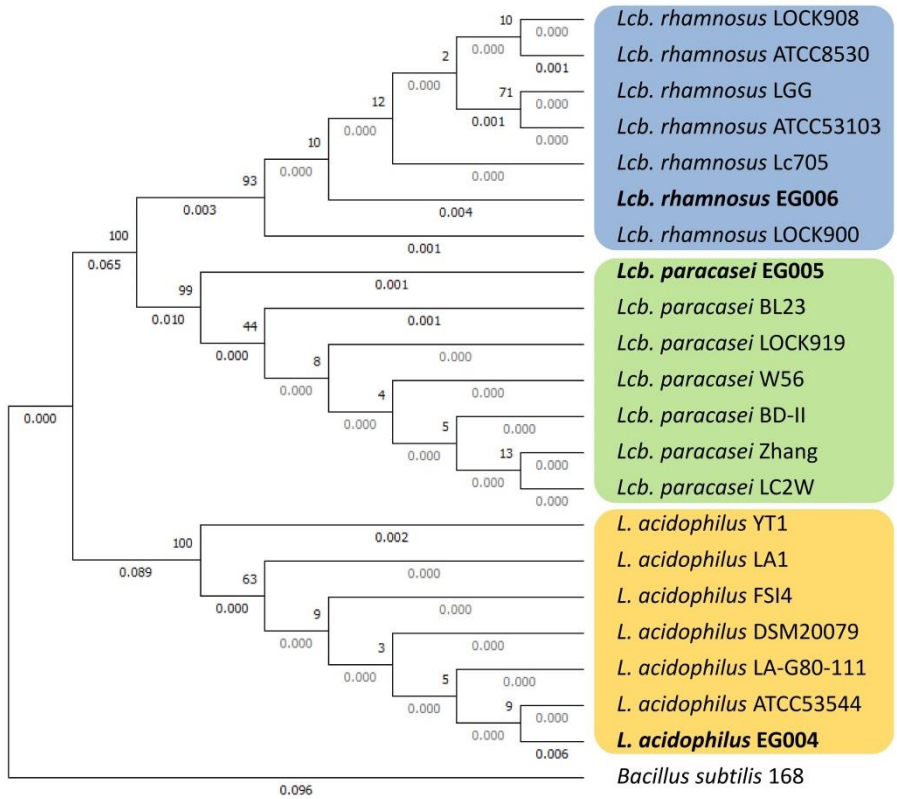
Among detected species with the ratio difference, several species were indicated as important factors in the research of brain disease. *Adlercreutzia equolifaciens* is equol (phytoestrogen) producing bacteria, which obstructs microglial function. In previous studies, a higher ratio of *A. equolifaciens* was found in the gut of patients with Alzheimer's disease and Autism spectrum disorder (Zhang, Ma et al. 2017, Laue, Korrick et al. 2020). In other studies, *Roseburia hominis* and *Bacteroides_F pectinophilus* were detected with a higher ratio in the patients with Alzheimer's disease than the normal persons (Haran, Bhattarai et al. 2019, Wang, Lei et al. 2019). When comparing gut microbiome between the Parkinson's disease group and normal group, *Soleaferrea massiliensis* was more frequently discovered in the patient group (Petrov, Alifirova et al. 2017). Interestingly, those strains

that showed a high ratio from the previous studies of brain disease patients were found to show a lower ratio in *L. acidophilus* group when compared to the control group (Figure 4F). Decreased bacterial ratio related to brain diseases seems to positively affect cognitive ability and we believe that it is due to *L. acidophilus* consumption. Although the specific mechanism cannot be estimated in this study, it seems to be influenced by the ingestion of *L. acidophilus* EG004. We hope that it will be a clue to unravel the role of *L. acidophilus* in the brain-gut axis in further studies.

In functional profiling analysis, we offered explainable factors for the microbial effect on the brain. Three KEGG categories were related to toxic chemical degradation: Dioxin degradation, Xylene degradation, and Caprolactam degradation (Figure 5B). Dioxin, a neurotoxin, can raise autism and neurodegenerative disease (Ames, Warner et al. 2018, Guo, Xie et al. 2018). Xylene inhibits normal protein synthesis of neuronal function and induces instability in the neuronal membrane. When it is inhaled, psychological deficits can be caused (Savolainen and Pfäffli 1980, Kandyala, Raghavendra et al. 2010). These chemicals are noxious to the brain, so activation of these chemical degradations would have diminished negative effects in *L. acidophilus* group. Besides, two KEGG categories related to the immune system were found. One of them is *Staphylococcus aureus* infection, which is known to cause brain abscess. Since there have been many studies demonstrating that *L. acidophilus* has antimicrobial activity against *S.*

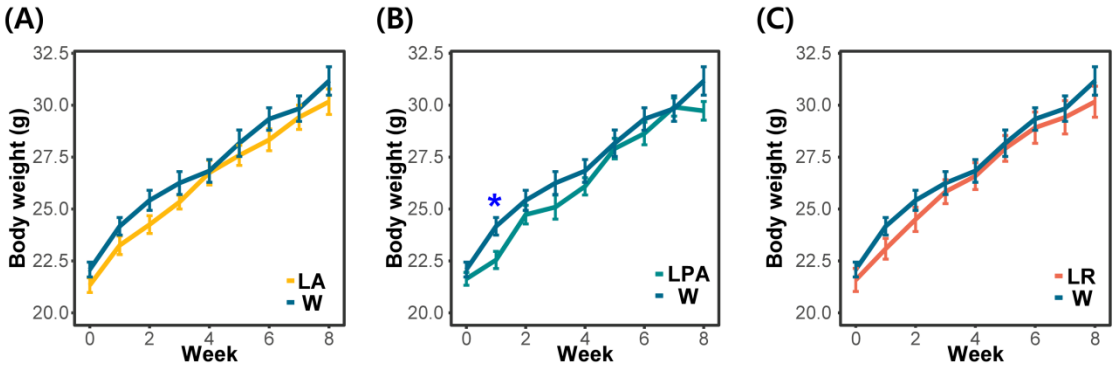
aureus, activation of this category is thought to be due to an increase in the amount of *L. acidophilus*. The function of renal cell carcinoma was predicted in the experimental group. As it involves not only tumor suppressor genes such as VHL, GH, and BHD, but also oncogenes such as MET and PRCC-TFE3, it seems to be necessary to confirm the exact mechanism and side effects.

The purpose of this study was to develop a new strain that has positive effects on brain function, which can be recognized through changes in cognitive processes. Also, we aimed to provide an underlying biological mechanism affecting the brain by the gut microbiome. It is necessary to measure metabolite changes in order to provide an understanding of the mechanism of altered cognitive ability. However, altered metabolite from animal body was not fully identified. To overcome this limitation, we conducted the metagenome analysis, correlation analysis between cognitive ability and gut microbiome, measurement of SCFA producing ability, and whole-genome comparison analysis. These analyses were not covered in the identification of a biological factor that caused improved cognitive ability, but presented a group of genes and mechanisms that can infer the process. Although we did not provide direct evidence of phenotype changes caused by probiotic ingestion, we hope that our findings will help infer the process of the brain-gut axis.



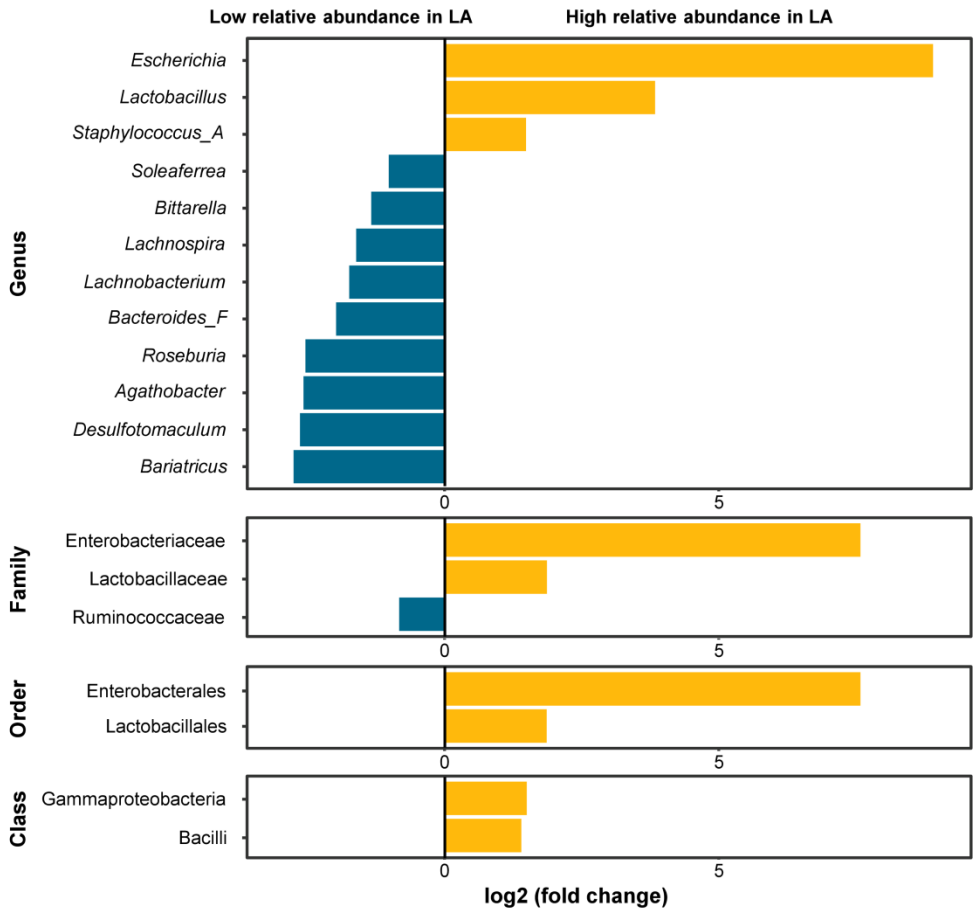
Additional file 16 - Figure S6-8. Phylogenetic tree using 16S rRNA of three probiotic strains

Phylogenetic tree by Maximum likelihood method based on 16S rRNA sequences for 14 *Lacticaseibacillus* genera and 7 *L. acidophilus* species including three strains in our study. Bootstrapping was conducted 1000 times, and the *Bacillus subtilis* 168 was used as an outgroup. The same species were enclosed in colored boxes, and three strains used in our study were presented in bold font.



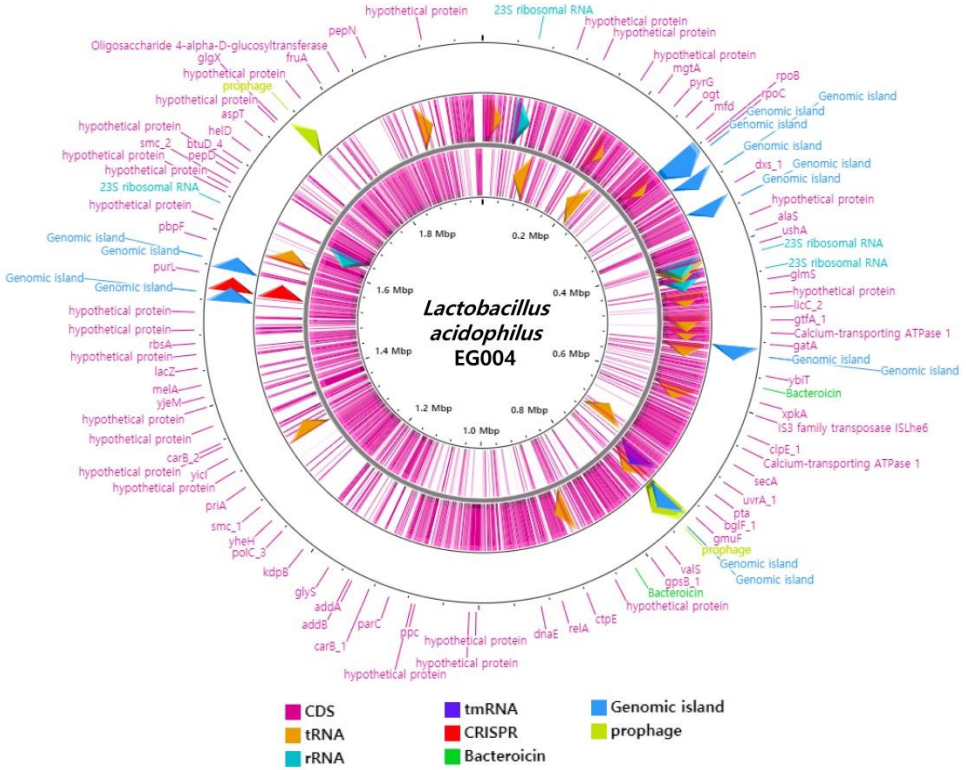
Additional file 17 - Figure S6-9. Animal body weight changes by week

We measured body weight of the mice every week. The average difference was found between the group fed *Lcb. paracasei* and control, but it was immediately recovered.



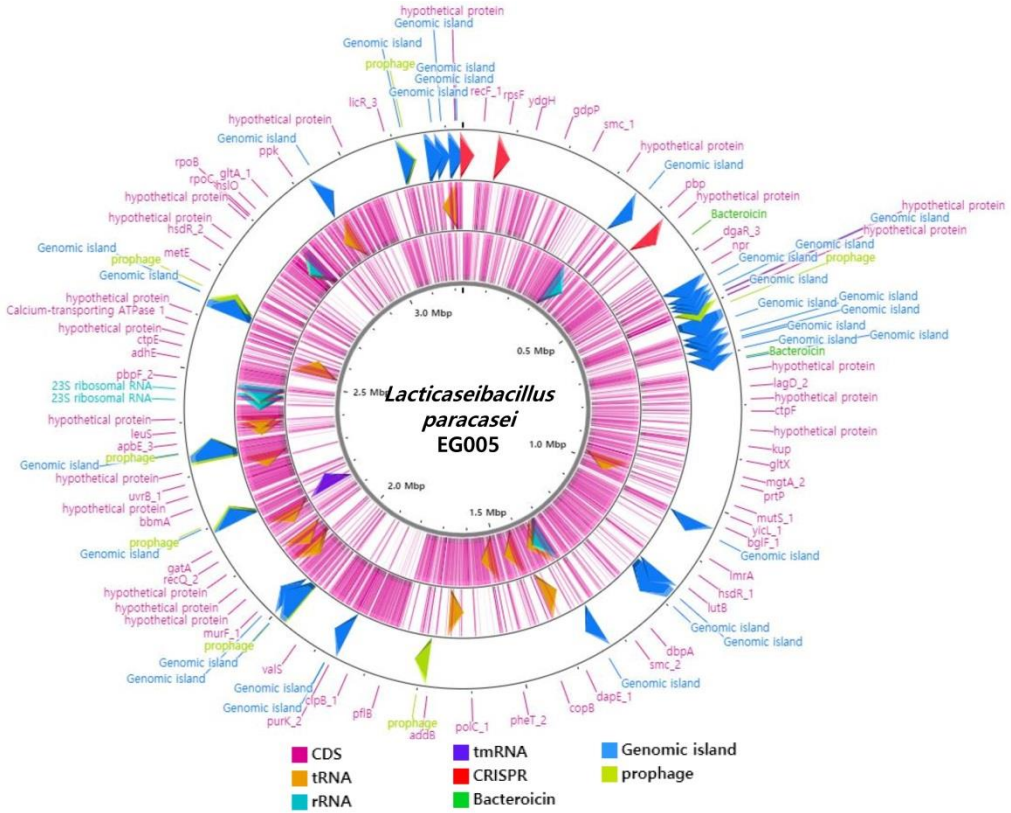
Additional file 18 - Figure S6-10. Comparison of microbial composition between the group fed *L. acidophilus* and control

We compared microbial composition of the group fed *L. acidophilus* and control at all taxonomic levels. The taxonomy found with high relative abundance in the group fed *L. acidophilus* showed yellow bar, while the other showed turquoise colored bar.



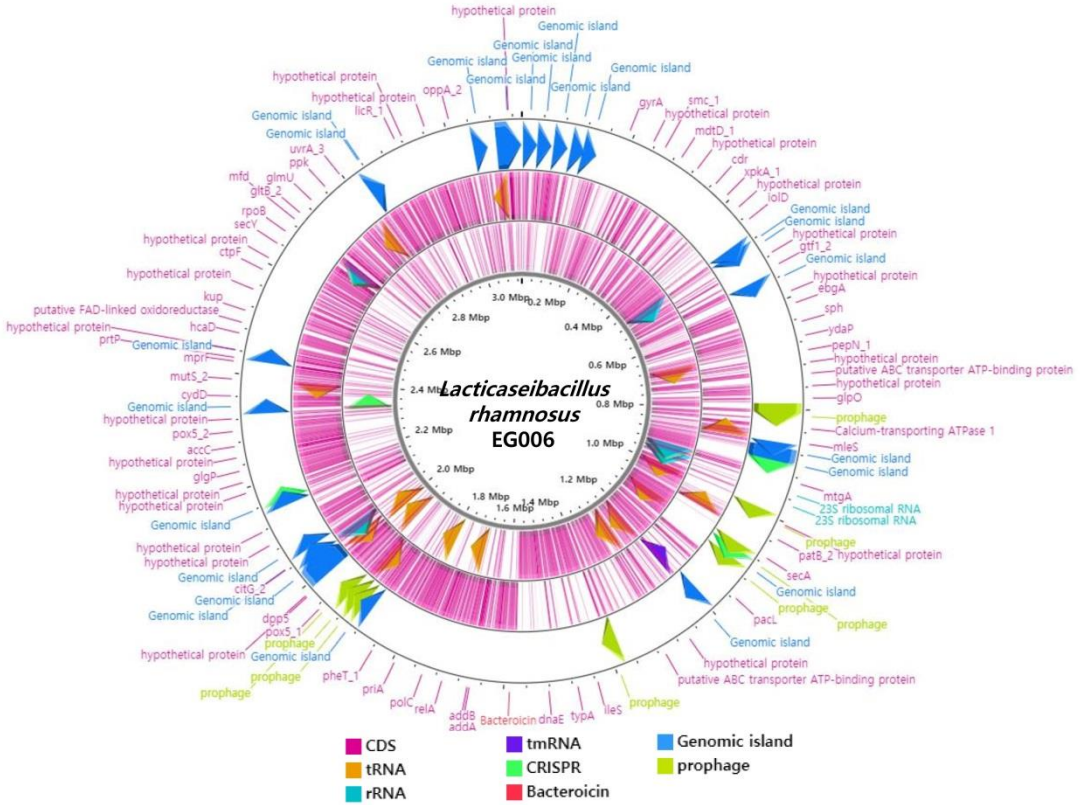
Additional file 19 - Figure S6-11. Circularized genome of *L. acidophilus* EG004

The genetic map represents main genetic factors of *L. acidophilus* EG004. There are three rings in generated, which indicates potential virulence factors (outermost ring), genes encoded on the forward strand, genes encoded on the backward stand (innermost ring).



Additional file 20 - Figure S6-12. Circularized genome of *Lcb. paracasei* EG005

This genetic map represents main genetic factors of *Lcb. paracasei* EG005. There are three rings in generated, which indicates potential virulence factors (outermost ring), genes encoded on the forward strand, genes encoded on the backward stand (innermost ring).



Additional file 21 - Figure S6-13. Circularized genome of *Lcb. rhamnosus* EG006

The physical map shows main genetic factors of *Lcb. rhamnosus* EG006. There are three rings in generated, which indicates potential virulence factors (outermost ring), genes encoded on the forward strand, genes encoded on the backward stand (innermost ring).

Additional file 22 - Table S5-1. Results of cognitive behavioral tests

Test	Assessment parameter	LA[a]	LPA[b]	LR[c]	Control
SA	Spontaneous alternation [*][%]	*66.08±6.11*	62.36±4.96	59.42±4.51	52.02±2.58
	Total arm entries	13.83±0.41	11.36±0.34	13.75±0.27	13.83±0.36
	Group ratio [d,*] [%]	*75.00*	63.64	58.33	66.67
NOR	Discrimination ratio [*] [%]	63.39±4.59	*64.53±6.06*	51.50±2.40	46.92±4.89
	Group ratio [e,*] [%]	*83.33*	72.73	41.67	60.00
	Time $_{Familiar}$ [f] (s)	4.10±0.76	2.75±0.68	4.83±0.92	4.32±0.81
	Time $_{Novel}$ [g] (s)	6.73±0.92	5.74±1.29	5.47±1.30	3.68±0.61
	Number $_{Familiar}$ [h]	5.17±0.65	3.00±0.66	5.92±0.72	5.50±0.95
	Number $_{Novel}$ [i]	7.75±0.91	5.45±0.92	7.50±1.59	5.40±0.83
PAT	Step latency at day 1 (s)	43.62±8.54	21.85±5.38	67.69±27.31	23.60±6.52
	Step latency at day 2 [*] (s)	*300±0.00*	281.20±14.05	292.59±7.41	193.79±41.97
FA	Forced alternation [*] [%]	28.92±5.72	20.63±5.04	*37.90±3.84*	23.33±6.03
	Total arm entries	18.20±3.09	10.71±0.81	17.36±1.16	15.42±1.73
	Group ratio [j,*] [%]	*83.33*	55.56	66.67	66.67

[a]: the group fed *L. acidophilus*, [b]: the group fed *Lcb. paracasei* [c]: the group fed *Lcb. rhamnosus*, [d]: The ratio of the mouse entered spontaneous alternation at first 3 entries, [e]: ratio of the mouse touched novel object at first, [f~i]: Time and number of touched Familiar or Novel objects, [j]: ratio of the mouse entered novel arm, and [*]: items for cognitive ability evaluation score. The highest value of cognitive ability indicators among the experimental groups was represented with red colored italic font. All values are shown as the mean ± SEM.

Additional file 23 - Table S5-3. SCFA identification in bacterial culture

Strain	Ret.Time (min)	Peak Name	Height (µRIU)	Area (µRIU*min)	Rel.Area (%)	Amount (mg/L)	Rel.Amount (%)
EG004	15.66	Lactic acid	72.540	27.036	81.58	13493.703	74.55
	18.69	Acetic acid	14.314	6.103	18.42	4605.749	25.45
	Total		86.855	33.139	100.00	18099.452	100.00
EG005	15.65	Lactic acid	80.019	29.558	83.96	14752.368	77.58
	18.68	Acetic acid	13.566	5.649	16.04	4262.900	22.42
	Total		93.585	35.207	100.00	19015.268	100.00
EG006	15.65	Lactic acid	50.834	18.616	77.89	9291.048	69.96
	18.68	Acetic acid	12.576	5.286	22.11	3989.009	30.04
	Total		63.410	23.901	100.00	13280.058	100.00

Short-chain fatty acids were measured by HPLC analysis. Broth media cultured for 24 hours were used for the analysis.

Additional file 24 - Table S5-4. Cognitive ability assessment score

Test	Evaluation item	LA[a]	LPA[b]	LR[c]	Control
SA	Spontaneous alternation [%]	4	3	2	1
	Group ratio [d] [%]	4	2	1	3
NOR	Discrimination ratio [%]	3	4	2	1
	Group ratio [e] [%]	4	3	1	2
PAT	Step latency at day 2 (s)	4	2	3	1
	Forced alternation [%]	3	1	4	2
FA	Group ratio [i] [%]	4	1	2	2
	Total	26	16	15	12

[a]: the group fed *L. acidophilus*, [b]: the group fed *Lcb. paracasei* [c]: the group fed *Lcb. rhamnosus*, [d]: The ratio of the mouse entered spontaneous alternation at first 3 entries, [e]: ratio of the mouse touched novel object at first, and [i]: ratio of the mouse entered novel arm. Scores of each cognitive ability assessment were given in ascending order of ranking (1-4 points).

General discussion

Sequencing technology has advanced dramatically over the past two decades. With the development of sequencing technology, methods for analyzing the characteristics of microbial genomes have also been diversified. Based on these technological advances, a lot of microbial genome data has been accumulated, and many researchers have tried to uncover the biological mechanism and evolutionary characteristics of microorganisms. The genome of LAB is small-scale and a monoploid, and it contains all the genetic information necessary for survival. Therefore, decoding the genome of microorganisms will not only understand the basic genetic mechanism of organisms but will also help understand the human genome. This thesis was conducted with the goal of identifying the level at which a microorganism can be understood and presenting its characteristics by applying the current sequencing technology.

For a full understanding of the microbial genome, genome analysis at various levels was performed. First, for the building of a complete genome, whole-genome sequencing was performed using PacBio and Nanopore technology, and metagenome sequencing was performed based on Nanopore technology. To understand a whole-genome sequence of microbial, analyses including genome annotation, functional protein categorization, protein 3D modeling, and gene detection were performed. To compare multiple

genomes, comparative analysis such as phylogenetic tree preparation, dN/dS analysis, codon usage comparison, resequencing, and pan-genome analysis was carried out according to the analysis purpose. Metagenome sequencing, biological diversity, taxonomy comparison, and functional profiling were performed to understand the microbial community. It was possible to identify gene contents, detect specific SNP, and compare ratios in the community. The identification of gene contents suggested the gene group possessed by one individual and further suggested the specificity shown in the individual by comparing multiple genomes at the gene level. Specific SNP detection presented a mutation of one nucleotide in the size of 3 MB. This included findings in the genic region as well as in the intergenic region that did not encode the gene. This suggests that the current genome analysis technology can detect a whole range of single mutations in the genome. The detection of gene contents and SNPs through whole-genome sequencing demonstrated that it was possible to construct a genome without an existing reference and to detect genetic content and single nucleotides. There are still problems that gene annotation has limitations depending on the capacity of the database used as a reference, and the detection of single nucleotides is sensitive to sequencing errors. However, these problems are expected to be solved soon with the development of sequencing technology and the accumulation of sequenced data. Finally, metagenome sequencing identified the ratio in the community of microorganisms using the rRNA gene.

Furthermore, a prediction of the expected functionality of the identified microbial community was performed. This shows that genome analysis technology can be applied to a new area beyond the existing whole-genome sequencing of single individuals.

To understand microorganism and their genome was not only applied the analysis method at various levels but also performed the analysis from more diverse viewpoints. It was mentioned earlier that the understanding of the microbial genome is the basis for understanding complex organisms such as a human. Approaches to understanding living organisms are forward genetics and reverse genetics. Forward genetics is a method to find individuals with different phenotypes and identify the genetic factors that cause the difference. The research conducted in Chapter 4 and Chapter 5 of this dissertation took this approach. In contrast, reverse genetics is a method of observing the phenotype that appears by inducing a mutation in a gene to investigate the function. Chapter 2 and Chapter 3 of this thesis performed genome analysis, which is the basis of reverse genetics. Although this dissertation did not confirm the suggested hypothesis by phenotype changing through genome editing technology such as CRISPR. However, these two-way approaches to the microorganism will clarify the relationship between phenotype and genotype.

However, several limitations remain. First, the experimental confirmation for the proposed genetic factors has not been done. Candidate factors were presented by comparative analysis with different phenotypes, but the relationship between the genetic factors and phenotypes could not be fully interpreted. This had limitations in the experimental design and the range of experiments that could be performed in the laboratory. Second, the series of studies conducted in this thesis only applied the well-researched method to a new probiotic strain. Although the reliability of the analysis was increased by using the previously established method, it was passive in expanding the scope of use of genome analysis to understand microorganisms. However, it is suggested that the significance of this study is to decipher the novel full-length genome and to suggest a new function of lactic acid bacteria.

This thesis was carried out to understand the LAB or the microbial complex ecosystem through various genome analyses. Decoding the genome of LAB is very important because it not only has useful functions for humans but also contains all the information necessary for survival in a short genome. To understand it, various genome analysis methods such as dN/dS analysis and metagenomic analysis were used, and various statistical methods were also supported. It is hoped that this study will be a step toward a complete understanding of the basic mechanisms of living organisms and genome sequencing.

References

Acman, M., et al. (2020). "Large-scale network analysis captures biological features of bacterial plasmids." Nature communications **11**(1): 1-11.

Ahmadian, A., et al. (2006). "Pyrosequencing: history, biochemistry and future." Clinica chimica acta **363**(1-2): 83-94.

Akhter, S., et al. (2012). "PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies." Nucleic acids research **40**(16): e126-e126.

Almpanis, A., et al. (2018). "Correlation between bacterial G+ C content, genome size and the G+ C content of associated plasmids and bacteriophages." **4**(4).

Altermann, E., et al. (2005). "Complete genome sequence of the probiotic lactic acid bacterium Lactobacillus acidophilus NCFM." Proceedings of the National Academy of Sciences **102**(11): 3906-3912.

Ames, J., et al. (2018). "Neurocognitive and physical functioning in the Seveso Women's Health Study." Environmental research **162**: 55-62.

Arndt, D., et al. (2016). "PHASTER: a better, faster version of the PHAST phage search tool." Nucleic acids research **44**(W1): W16-W21.

Aziz, R. K., et al. (2008). "The RAST Server: rapid annotations using subsystems technology." BMC genomics **9**(1): 75.

Aziz, R. K., et al. (2008). "The RAST Server: rapid annotations using subsystems technology." **9**(1): 75.

Bairoch, A. and B. J. N. a. r. Boeckmann (1991). "The SWISS-PROT protein sequence data bank." Nucleic acids research **19**(Suppl): 2247.

Barichello, T., et al. (2015). "Sodium butyrate prevents memory impairment by re-establishing BDNF and GDNF expression in experimental pneumococcal meningitis." Molecular neurobiology **52**(1): 734-740.

Basak, S., et al. (2010). "Genomic adaptation of prokaryotic organisms at high temperature." Bioinformation **4**(8): 352.

Beaty, T. H., et al. (2011). "Evidence for gene-environment interaction in a genome wide study of nonsyndromic cleft palate." Genetic epidemiology **35**(6): 469-478.

Belhadj Slimen, I., et al. (2016). "Heat stress effects on livestock: molecular, cellular and metabolic aspects, a review." Journal of animal physiology **100**(3): 401-412.

Bernal-Cabas, M., et al. (2015). "The Cpx envelope stress response modifies peptidoglycan cross-linking via the L, D-transpeptidase LdtD and the novel protein YgaU." Journal of bacteriology **197**(3): 603-614.

Bertelli, C., et al. (2017). "IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets." Nucleic acids research **45**(W1): W30-W35.

Bertrand, J. A., et al. (1997). "Crystal structure of UDP-N-acetylmuramoyl-L-alanine: D-glutamate ligase from Escherichia coli." The EMBO journal **16**(12): 3416-3425.

Bertuccini, L., et al. (2017). "Effects of Lactobacillus rhamnosus and Lactobacillus acidophilus on bacterial vaginal pathogens." International journal of immunopathology **30**(2): 163-167.

Bohlin, J., et al. (2010). "Analysis of intra-genomic GC content homogeneity within prokaryotes." BMC genomics **11**(1): 1-8.

Bohlin, J., et al. (2012). "Relative entropy differences in bacterial chromosomes, plasmids, phages and genomic islands." BMC genomics **13**(1): 1-12.

Boney, J., et al. (2018). "The effects of steam conditioning and antimicrobial inclusion on feed manufacturing and inactivation of Enterococcus faecium, a Salmonella surrogate." Journal of Applied Poultry Research **27**(4): 472-482.

Bose, T., et al. (2015). "COGNIZER: a framework for functional annotation of metagenomic datasets." PloS one **10**(11): e0142102.

Bourassa, M. W., et al. (2016). "Butyrate, neuroepigenetics and the gut microbiome: can a high fiber diet improve brain health?" Neuroscience letters **625**: 56-63.

Brammer, L. B., et al. (2015). "Loss of a Functionally and Structurally Distinct ld-Transpeptidase, LdtMt5, Compromises Cell Wall Integrity in Mycobacterium tuberculosis." The Journal of biological chemistry **290**(42): 25670-25685.

Cao, Y.-N., et al. (2018). "Lactobacillus acidophilus and Bifidobacterium longum supernatants upregulate the serotonin transporter expression in intestinal epithelial cells." Saudi journal of gastroenterology **24**(1): 59.

Carver, T., et al. (2009). "DNAPlotter: circular and linear interactive genome visualization." Bioinformatics **25**(1): 119-120.

Castresana, J. (2000). "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis." Molecular biology and evolution **17**(4): 540-552.

Chahuki, F. F., et al. (2019). "Hyaluronic acid production enhancement via genetically modification and culture medium optimization in Lactobacillus acidophilus." International journal of biological macromolecules **121**: 870-881.

Chaillou, S., et al. (2005). "The complete genome sequence of the meat-borne lactic acid bacterium Lactobacillus sakei 23K." Nature biotechnology **23**(12): 1527-1533.

Chen, F., et al. (2013). "The history and advances of reversible terminators used in new generations of sequencing technology." Genomics, Proteomics Bioinformatics **11**(1): 34-40.

Chen, M.-J., et al. (2017). "Effects of heat, cold, acid and bile salt adaptations on the stress tolerance and protein expression of kefir-isolated probiotic Lactobacillus kefiranofaciens M1." Food microbiology **66**: 20-27.

Chen, W.-H., et al. (2016). "Energy efficiency trade-offs drive nucleotide usage in transcribed regions." Nature communications **7**(1): 1-10.

Chen, Y., et al. (2014). "edgeR: differential expression analysis of digital gene expression data User's Guide." Bioconductor User's Guide.

Chervaux, C., et al. (2000). "Physiological study of Lactobacillus delbrueckii subsp. bulgaricus strains in a novel chemically defined medium." Applied Environmental Microbiology
66(12): 5306-5311.

Chin, C.-S., et al. (2013). "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data." Nature methods **10**(6): 563-569.

Choi, S., et al. (2006). "Effects of Lactobacillus strains on cancer cell proliferation and oxidative stress in vitro." Letters in Applied Microbiology **42**(5): 452-458.

Chou, L.-S. and B. J. J. o. D. S. Weimer (1999). "Isolation and characterization of acid-and bile-tolerant isolates from strains of Lactobacillus acidophilus." Journal of Dairy Science **82**(1): 23-31.

Cingolani, P., et al. (2012). "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3." Fly **6**(2): 80-92.

Coconnier, M.-H., et al. (1997). "Antibacterial effect of the adhering human Lactobacillus acidophilus strain LB." **41**(5): 1046-1052.

Corpuz, H. M., et al. (2018). "Long-term diet supplementation with Lactobacillus paracasei K71 prevents age-related cognitive decline in senescence-accelerated mouse prone 8." **10**(6): 762.

Crick, F. (1970). "Central dogma of molecular biology." Nature **227**(5258): 561-563.

Daeschel, M., et al. (1990). "Bacteriocidal activity ofLactobacillus plantarum C-11." Food Microbiology **7**(2): 91-98.

Danilenko, V. N., et al. (2020). "The use of a pharmabiotic based on the Lactobacillus fermentum U-21 strain to modulate the neurodegenerative process in an experimental model of Parkinson disease." Annals of Clinical Experimental Neurology **14**(1).

Dash, S. K. J. A. F. I. H. T. (2004). "Review of scientific evidence for efficacy of Lactobacillus acidophilus DDS-1 as a probiotic strain." AGRO FOOD INDUSTRY HI TECH **15**(5): 23-26.

De Angelis, M., et al. (2016). "Functional proteomics within the genus Lactobacillus." Proteomics **16**(6): 946-962.

de Jong, A., et al. (2006). "BAGEL: a web-based bacteriocin genome mining tool." Nucleic acids research **34**(suppl_2): W273-W279.

Dennis, G., et al. (2003). "DAVID: database for annotation, visualization, and integrated discovery." Genome biology **4**(9): 1-11.

Desiere, F., et al. (2001). "Comparative genomics reveals close genetic relationships between phages from dairy bacteria and pathogenic streptococci: evolutionary implications for prophage-host interactions." Virology **288**(2): 325-341.

Dobrindt, U., et al. (2004). "Genomic islands in pathogenic and environmental microorganisms." Nature Reviews Microbiology **2**(5): 414-424.

Dobson, A., et al. (2012). "Bacteriocin production: a probiotic trait?" Applied environmental microbiology **78**(1): 1-6.

Douglas, G. M., et al. (2020). "PICRUSt2: An improved and customizable approach for metagenome inference." BioRxiv: 672295.

Dragosits, M. and D. J. M. c. f. Mattanovich (2013). "Adaptive laboratory evolution–principles and applications for biotechnology." Microbial cell factories **12**(1): 1-17.

Du, M.-Z., et al. (2018). "The GC content as a main factor shaping the amino acid usage during bacterial evolution process." Frontiers in microbiology **9**: 2948.

Duar, R. M., et al. (2017). "Lifestyles in transition: evolution and natural history of the genus Lactobacillus." FEMS microbiology reviews **41**(Supp_1): S27-S48.

Dumas, A., et al. (2018). "The role of the lung microbiota and the gut–lung axis in respiratory infectious diseases." Cellular microbiology **20**(12): e12966.

Emms, D. M. and S. J. G. b. Kelly (2019). "OrthoFinder: phylogenetic orthology inference for comparative genomics." Genome biology **20**(1): 1-14.

Felis, G. E. and F. J. C. i. i. i. m. Dellaglio (2007). "Taxonomy of lactobacilli and bifidobacteria." Current issues in intestinal microbiology **8**(2): 44.

Ferrando, V., et al. (2015). "Resistance of functional Lactobacillus plantarum strains against food stress conditions." Food microbiology **48**: 63-71.

Fetissov, S. O., et al. (2019). "Neuropeptides in the microbiota-brain axis and feeding behavior in autism spectrum disorder." Nutrition **61**: 43-48.

Firtina, C., et al. (2020). "Apollo: a sequencing-technology-independent, scalable and accurate assembly polishing algorithm." Bioinformatics **36**(12): 3669-3679.

Foysal, M. J., et al. (2020). "Lactobacillus acidophilus and L. plantarum improve health status, modulate gut microbiota and innate immune response of marron (Cherax cainii)." Scientific reports **10**(1): 1-13.

Fuentes, M. C., et al. (2013). "Cholesterol-lowering efficacy of Lactobacillus plantarum CECT 7527, 7528 and 7529 in hypercholesterolaemic adults." British Journal of Nutrition **109**(10): 1866-1872.

Gálvez, A., et al. (2007). "Bacteriocin-based strategies for food biopreservation." International journal of food microbiology **120**(1-2): 51-70.

Ghaisas, S., et al. (2016). "Gut microbiome in health and disease: Linking the microbiome–gut–brain axis and environmental factors in the pathogenesis of systemic and neurodegenerative diseases." Pharmacology therapeutics **158**: 52-62.

Ghouri, Y. A., et al. (2014). "Systematic review of randomized controlled trials of probiotics, prebiotics, and synbiotics in inflammatory bowel disease." Clinical experimental gastroenterology **7**: 473.

Glaasker, E., et al. (1998). "Physiological response of Lactobacillus plantarum to salt and nonelectrolyte stress." Journal of bacteriology **180**(17): 4718-4723.

Goodwin, S., et al. (2016). "Coming of age: ten years of next-generation sequencing technologies." Nature Reviews Genetics **17**(6): 333-351.

Grada, A. and K. J. T. J. o. i. d. Weinbrecht (2013). "Next-generation sequencing: methodology and application." The Journal of investigative dermatology **133**(8): e11.

Grant, J. R. and P. J. N. a. r. Stothard (2008). "The CGView Server: a comparative genomics tool for circular genomes." **36**(suppl_2): W181-W184.

Grissa, I., et al. (2007). "CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats." Nucleic acids research **35**(suppl_2): W52-W57.

Grosu-Tudor, S.-S., et al. (2016). "S-layer production by Lactobacillus acidophilus IBB 801 under environmental stress conditions." Applied microbiology biotechnology **100**(10): 4573-4583.

Güvenç, I. A., et al. (2016). "Do probiotics have a role in the treatment of allergic rhinitis? A comprehensive systematic review and Metaanalysis." American journal of rhinology allergy **30**(5): e157-e175.

Guo, Z., et al. (2018). "Dioxins as potential risk factors for autism spectrum disorder." Environment international **121**: 906-915.

Gupta, R., et al. (2010). "The Mycobacterium tuberculosis protein Ldt Mt2 is a nonclassical transpeptidase required for virulence and resistance to amoxicillin." Nature medicine **16**(4): 466-469.

Gupta, S. K., et al. (2014). "ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes." Antimicrobial agents **58**(1): 212-220.

Haldar, S., et al. (2011). "Effects of yeast (Saccharomyces cerevisiae) and yeast protein concentrate on production performance of broiler chickens exposed to heat stress and challenged with Salmonella enteritidis." Animal Feed Science Technology **168**(1-2): 61-71.

Haran, J. P., et al. (2019). "Alzheimer's disease microbiome is associated with dysregulation of the anti-inflammatory P-glycoprotein pathway." MBio **10**(3).

Hayek, N. J. F. i. M. (2013). "Lateral transfer and GC content of bacterial resistance genes." Frontiers in microbiology **4**: 41.

He, F., et al. (2014). "Glatiramer acetate reverses cognitive deficits from cranial-irradiated rat by inducing hippocampal neurogenesis." Journal of neuroimmunology **271**(1-2): 1-7.

Hegarty, J. W., et al. (2016). "Bacteriocin production: a relatively unharnessed probiotic trait?" FResearch **5**.

Hertz, L., et al. (2015). "Effects of ketone bodies in Alzheimer's disease in relation to neural hypometabolism, β-amyloid toxicity, and astrocyte function." Journal of neurochemistry **134**(1): 7-20.

Hicke, L. and R. Dunn (2003). "Regulation of membrane protein transport by ubiquitin and ubiquitin-binding proteins." Annual review of cell and developmental biology **19**(1): 141-172.

Higgins, C. F. (1992). "ABC transporters: from microorganisms to man." Annual review of cell biology **8**(1): 67-113.

Higgs, P. G., et al. (2008). "Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage." Molecular biology evolution **25**(11): 2279-2291.

Hildebrand, F., et al. (2010). "Evidence of selection upon genomic GC-content in bacteria." PLoS genetics **6**(9): e1001107.

Hill-Burns, E. M., et al. (2017). "Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome." Movement Disorders **32**(5): 739-749.

Hirota, R., et al. (2010). "Bacterial phosphate metabolism and its application to phosphorus recovery and industrial bioprocesses." Journal of bioscience and bioengineering **109**(5): 423-432.

Hoang, B. X., et al. (2010). "Lactobacillus rhamnosus cell lysate in the management of resistant childhood atopic eczema." Inflammation Allergy-Drug Targets **9**(3): 192-196.

Hogue, C. W. (1997). "Cn3D: a new generation of three-dimensional molecular structure viewer." Trends in biochemical sciences **22**(8): 314-316.

Hove-Jensen, B., et al. (2014). "Utilization of glyphosate as phosphate source: biochemistry and genetics of bacterial carbon-phosphorus lyase." Microbiology and Molecular Biology Reviews **78**(1): 176-197.

Hsieh, Y.-J. and B. L. Wanner (2010). "Global regulation by the seven-component P i signaling system." Current opinion in microbiology **13**(2): 198-203.

Hu, K., et al. (2021). "MultiNanopolish: refined grouping method for reducing redundant calculations in Nanopolish." Bioinformatics **37**(17): 2757-2760.

Huang, J., et al. (2005). "The evolution of microbial phosphonate degradative pathways." Journal of molecular evolution **61**(5): 682-690.

Huang, S.-Y., et al. (2018). "Lactobacillus paracasei PS23 delays progression of age-related cognitive decline in senescence accelerated mouse prone 8 (SAMP8) mice." Nutrients **10**(7): 894.

Huerta-Cepas, J., et al. (2017). "Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper." Molecular biology evolution **34**(8): 2115-2122.

Hugonnet, J.-E., et al. (2016). "Factors essential for L, D-transpeptidase-mediated peptidoglycan cross-linking and β-lactam resistance in Escherichia coli." Elife **5**: e19469.

Hunt, M., et al. (2015). "Circlator: automated circularization of genome assemblies using long sequencing reads." **16**(1): 294.

Hurst, L. D. and A. R. J. P. o. t. R. S. o. L. S. B. B. S. Merchant (2001). "High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes." Proceedings of the Royal Society of London **268**(1466): 493-497.

Hutkins, R. W., et al. (1987). "Betaine transport imparts osmotolerance on a strain of Lactobacillus acidophilus." Applied environmental microbiology **53**(10): 2275-2281.

Im, A.-R., et al. (2018). "Skin moisturizing and antiphotodamage effects of tyndallized Lactobacillus acidophilus IDCC 3302." Journal of medicinal food **21**(10): 1016-1023.

Jain, V., et al. (2006). "ppGpp: stringent response and survival." Journal of microbiology **44**(1): 1-10.

Jeon, S., et al. (2021). "Complete Genome Sequence of the Newly Developed Lactobacillus acidophilus Strain With Improved Thermal Adaptability." Frontiers in microbiology: 2771.

Jia, B., et al. (2016). "CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database." Nucleic acids research: gkw1004.

Joensen, K. G., et al. (2014). "Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli." Journal of clinical microbiology **52**(5): 1501-1510.

Johansson, M., et al. (1993). "Administration of different Lactobacillus strains in fermented oatmeal soup: in vivo colonization of human intestinal mucosa and effect on the indigenous flora." Applied and Environmental Microbiology **59**(1): 15-20.

Johnson, M., et al. (2008). "NCBI BLAST: a better web interface." **36**(suppl_2): W5-W9.

Jones, F. and K. J. P. s. Richardson (2004). "Salmonella in commercially manufactured feeds." **83**(3): 384-391.

Jordan, K. and T. J. L. i. A. M. Cogan (1999). "Heat resistance of Lactobacillus spp. isolated from Cheddar cheese." Letters in Applied Microbiology **29**(2): 136-140.

Juhas, M., et al. (2009). "Genomic islands: tools of bacterial horizontal gene transfer and evolution." FEMS microbiology reviews **33**(2): 376-393.

Jung, W. Y., et al. (2016). "Functional characterization of bacterial communities responsible for fermentation of doenjang: a traditional Korean fermented soybean paste." Frontiers in microbiology **7**: 827.

Kandyala, R., et al. (2010). "Xylene: An overview of its health hazards and preventive measures." Journal of oral maxillofacial pathology: JOMFP **14**(1): 1.

Kapli, P., et al. (2020). "Phylogenetic tree building in the genomic age." Nature Reviews Genetics **21**(7): 428-444.

Karl, D. M. (2000). "Aquatic ecology: Phosphorus, the staff of life." Nature **406**(6791): 31-33.

Karst, S. M., et al. (2021). "High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing." Nature methods **18**(2): 165-169.

Kaur, H., et al. (2019). "Tryptophan metabolism by gut microbiome and gut-brain-axis: an in silico analysis." Frontiers in neuroscience **13**: 1365.

Kelly, J. R., et al. (2017). "Lost in translation? The potential psychobiotic Lactobacillus rhamnosus (JB-1) fails to modulate stress or cognitive performance in healthy male subjects." Brain, behavior, immunity **61**: 50-59.

Khaleghi, M., et al. (2010). "Assessment of bile salt effects on S-layer production, slp gene expression and, some physicochemical properties of Lactobacillus acidophilus ATCC 4356." **20**(4): 749-756.

Khoury, M. J. and S. Wacholder (2009). "Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies—challenges and opportunities." American journal of epidemiology **169**(2): 227-230.

Kim, K. M., et al. (2008). "An approach of orthology detection from homologous sequences under minimum evolution." Nucleic acids research **36**(17): e110-e110.

Kim, W. S., et al. (2001). "Assessment of stress response of the probiotic Lactobacillus acidophilus." Current Microbiology **43**(5): 346-350.

Kleerebezem, M., et al. (2003). "Complete genome sequence of Lactobacillus plantarum WCFS1." Proceedings of the National Academy of Sciences **100**(4): 1990-1995.

Klein, A., et al. (2008). "Lactobacillus acidophilus 74-2 and Bifidobacterium animalis subsp lactis DGCC 420 modulate unspecific cellular immune response in healthy adults." European journal of clinical nutrition **62**(5): 584-593.

Klein, P., et al. (2010). "Ketogenic diet treatment in adults with refractory epilepsy." **19**(4): 575-579.

Kogay, R., et al. (2020). "Selection for reducing energy cost of protein production drives the GC content and amino acid composition bias in gene transfer agents." **11**(4): e01206-01220.

Kolmogorov, M., et al. (2019). "Assembly of long, error-prone reads using repeat graphs." Nature biotechnology **37**(5): 540-546.

Komatsuzaki, N., et al. (2005). "Production of γ-aminobutyric acid (GABA) by Lactobacillus paracasei isolated from traditional fermented foods." Food microbiology **22**(6): 497-504.

Koren, S., et al. (2017). "Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation." Genome research **27**(5): 722-736.

Kryazhimskiy, S. and J. B. J. P. g. Plotkin (2008). "The population genetics of dN/dS." PLoS genetics **4**(12): e1000304.

Kuipers, O. P., et al. (1998). "Quorum sensing-controlled gene expression in lactic acid bacteria." Journal of Biotechnology **64**(1): 15-21.

Kulkarni, S., et al. (2018). "Adaptation of Lactobacillus acidophilus to thermal stress yields a thermotolerant variant which also exhibits improved survival at pH 2." **10**(4): 717-727.

Kumar, S., et al. (2018). "MEGA X: molecular evolutionary genetics analysis across computing platforms." **35**(6): 1547.

Kumar, S., et al. (2016). "MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets." Molecular biology and evolution: msw054.

Kurtz, S., et al. (2004). "Versatile and open software for comparing large genomes." Genome biology **5**(2): R12.

Kwak, S.-H., et al. (2014). "Cancer preventive potential of kimchi lactic acid bacteria (Weissella cibaria, Lactobacillus plantarum)." Journal of cancer prevention **19**(4): 253-258.

Lagesen, K., et al. (2007). "RNAmmer: consistent and rapid annotation of ribosomal RNA genes." **35**(9): 3100-3108.

Lan, R., et al. (2017). "Effects of Lactobacillus acidophilus supplementation on growth performance, nutrient digestibility, fecal microbial and noxious gas emission in weaning pigs." Journal of the Science of Food Agriculture **97**(4): 1310-1315.

Langille, M. G. and F. S. Brinkman (2009). "IslandViewer: an integrated interface for computational identification and visualization of genomic islands." Bioinformatics **25**(5): 664-665.

Larkin, M. A., et al. (2007). "Clustal W and Clustal X version 2.0." Bioinformatics **23**(21): 2947-2948.

Laue, H. E., et al. (2020). "Prospective associations of the infant gut microbiome and microbial function with social behaviors related to autism at age 3 years." Scientific reports **10**(1): 1-11.

Lee, H., et al. (2011). "Functional properties of Lactobacillus strains isolated from kimchi." International journal of food microbiology **145**(1): 155-161.

Leenay, R. T., et al. (2019). "Genome editing with CRISPR-Cas9 in Lactobacillus plantarum revealed that editing outcomes can vary across strains and between methods." Biotechnology journal **14**(3): 1700583.

Lehnert, H., et al. (1993). "Amino acid control of neurotransmitter synthesis and release: physiological and clinical implications." Psychotherapy psychosomatics **60**(1): 18-32.

Li, H., et al. (2009). "The sequence alignment/map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Li, J.-s., et al. (2011). "Transcriptome analysis of adaptive heat shock response of Streptococcus thermophilus." **6**(10): e25777.

Li, J., et al. (2018). "VRprofile: gene-cluster-detection-based profiling of virulence and antibiotic resistance traits encoded within genome sequences of pathogenic bacteria." Briefings in Bioinformatics **19**(4): 566-574.

Li, S., et al. (2012). "Antioxidant activity of Lactobacillus plantarum strains isolated from traditional Chinese fermented foods." Food chemistry **135**(3): 1914-1919.

Lim, S.-D., et al. (2009). "Physiological characteristics and GABA production of Lactobacillus acidophilus RMK567 isolated from raw milk." Food Science of Animal Resources **29**(1): 15-23.

Lim, S.-D., et al. (2009). "GABA productivity in yoghurt fermented by freeze dried culture preparations of Lactobacillus acidophilus RMK567." Food Science of Animal Resources **29**(4): 437-444.

Lind, P. A. and D. I. J. P. o. t. N. A. o. S. Andersson (2008). "Whole-genome mutational biases in bacteria." Proceedings of the National Academy of Sciences **105**(46): 17878-17883.

Liu, Y.-W., et al. (2018). "New perspectives of Lactobacillus plantarum as a probiotic: The gut-heart-brain axis." Journal of microbiology **56**(9): 601-613.

Löytynoja, A. and N. Goldman (2008). "Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis." Science **320**(5883): 1632-1635.

Lu, H., et al. (2016). "Oxford Nanopore MinION sequencing and genome assembly." Genomics, proteomics bioinformatics **14**(5): 265-279.

Makarova, K., et al. (2006). "Comparative genomics of the lactic acid bacteria." Proceedings of the National Academy of Sciences **103**(42): 15611-15616.

Mann, S. and Y.-P. P. J. G. Chen (2010). "Bacterial genomic G+ C composition-eliciting environmental adaptation." Genomics **95**(1): 7-15.

Mao, G., et al. (2019). "Depolymerized RG-I-enriched pectin from citrus segment membranes modulates gut microbiota, increases SCFA production, and promotes the growth of Bifidobacterium spp., Lactobacillus spp. and Faecalibaculum spp." Food function **10**(12): 7828-7843.

Martin-Gallausiaux, C., et al. (2021). "SCFA: mechanisms and functional importance in the gut." Proceedings of the Nutrition Society **80**(1): 37-49.

Martin, C. R., et al. (2018). "The brain-gut-microbiome axis." Cellular molecular gastroenterology hepatology
6(2): 133-148.

Matijašić, B. B., et al. (2016). "Effects of synbiotic fermented milk containing Lactobacillus acidophilus La-5 and Bifidobacterium animalis ssp. lactis BB-12 on the fecal microbiota of adults with irritable bowel syndrome: A randomized double-blind, placebo-controlled trial." Journal of Dairy Science **99**(7): 5008-5021.

Matsushita, K., et al. (2016). "Genomic analyses of thermotolerant microorganisms used for high-temperature fermentations." Bioscience, biotechnology, biochemistry **80**(4): 655-668.

Maxam, A. M. and W. J. P. o. t. N. A. o. S. Gilbert (1977). "A new method for sequencing DNA." Proceedings of the National Academy of Sciences **74**(2): 560-564.

Mayer, E. A., et al. (2014). "Brain–gut microbiome interactions and functional bowel disorders." Gastroenterology **146**(6): 1500-1512.

McDonald, L., et al. (1990). "Acid tolerance of Leuconostoc mesenteroides and Lactobacillus plantarum." Applied and Environmental Microbiology **56**(7): 2120-2124.

McMurdie, P. J. and S. J. P. o. Holmes (2013). "phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data." PloS one **8**(4): e61217.

McVey Neufeld, K.-A., et al. (2019). "Neurobehavioural effects of Lactobacillus rhamnosus GG alone and in combination with prebiotics polydextrose and galactooligosaccharide in male rats exposed to early-life stress." Nutritional Neuroscience **22**(6): 425-434.

Meena, S., et al. (2016). "Common mechanism of cross-resistance development in pathogenic bacteria bacillus cereus against alamethicin and pediocin involves alteration in lipid composition." Current Microbiology **73**(4): 534-541.

Menconi, A., et al. (2014). "Identification and characterization of lactic acid bacteria in a commercial probiotic culture." Bioscience of Microbiota, Food Health **33**(1): 25-30.

Mengin-Lecreulx, D., et al. (1999). "Expression of the Staphylococcus aureus UDP-N-acetylmuramoyl-L-alanyl-D-glutamate: L-lysine ligase in Escherichia coli and effects on peptidoglycan biosynthesis and cell growth." Journal of bacteriology **181**(19): 5909-5914.

Miettinen, M., et al. (1996). "Production of human tumor necrosis factor alpha, interleukin-6, and interleukin-10 is induced by lactic acid bacteria." Infection immunity **64**(12): 5403-5405.

Miller, M. B. and B. L. Bassler (2001). "Quorum sensing in bacteria." Annual Reviews in Microbiology **55**(1): 165-199.

Min, B., et al. (2020). "Complete Genomic Analysis of Enterococcus faecium Heat-Resistant Strain Developed by Two-Step Adaptation Laboratory Evolution Method." Frontiers in bioengineering and biotechnology **8**: 828.

Mitchell, D. J. B. and b. r. communications (2007). "GC content and genome length in Chargaff compliant genomes." Biochemical biophysical research communications **353**(1): 207-210.

Mohankumar, A. and N. J. I. J. o. B. Murugalatha (2011). "Characterization and antibacterial activity of bacteriocin producing Lactobacillus isolated from raw cattle milk sample." International Journal of Biology **3**(3): 128.

Newman, Z. R., et al. (2016). "Differences in codon bias and GC content contribute to the balanced expression of TLR7 and TLR9." Proceedings of the National Academy of Sciences **113**(10): E1362-E1371.

Nguyen, A., et al. (2015). "Isolation and characterization of B acillus subtilis CH 16 strain from chicken gastrointestinal tracts for use as a feed supplement to promote weight gain in broilers." Letters in Applied Microbiology **60**(6): 580-588.

Niedzielin, K., et al. (2001). "A controlled, double-blind, randomized study on the efficacy of Lactobacillus plantarum 299V in patients with irritable bowel syndrome." European journal of gastroenterology hepatology **13**(10): 1143-1147.

Norwitz, N. G., et al. (2019). "The mechanisms by which the ketone body D-β-hydroxybutyrate may improve the multiple cellular pathologies of Parkinson's disease." Frontiers in nutrition **6**: 63.

O'Mahony, S. M., et al. (2015). "Serotonin, tryptophan metabolism and the brain-gut-microbiome axis." Behavioural brain research **277**: 32-48.

Oh, N. S., et al. (2020). "Glycated milk protein fermented with Lactobacillus rhamnosus ameliorates the cognitive health of mice under mild-stress condition." **11**(6): 1643-1661.

Okpokwasili, G. and C. J. A. J. o. B. Nweke (2006). "Microbial growth and substrate utilization kinetics." African Journal of Biotechnology **5**(4): 305-317.

Ö sterlund, P., et al. (2007). "Lactobacillus supplementation for diarrhoea related to chemotherapy of colorectal cancer: a randomised study." British journal of cancer **97**(8): 1028-1034.

Palomino, M. M., et al. (2016). "Influence of osmotic stress on the profile and gene expression of surface layer proteins in Lactobacillus acidophilus ATCC 4356." **100**(19): 8475-8484.

Park, J.-S., et al. (2018). "Lactobacillus acidophilus improves intestinal inflammation in an acute colitis mouse model by regulation of Th17 and Treg cell balance and fibrosis development." Journal of medicinal food **21**(3): 215-224.

Parmley, J. L. and L. D. J. B. Hurst (2007). "How do synonymous mutations affect fitness?" Bioessays **29**(6): 515-519.

Peltier, J., et al. (2011). "Clostridium difficile has an original peptidoglycan structure with a high level of N-acetylglucosamine deacetylation and mainly 3-3 cross-links." Journal of Biological Chemistry **286**(33): 29053-29062.

Peran, L., et al. (2007). "A comparative study of the preventative effects exerted by three probiotics, Bifidobacterium lactis, Lactobacillus casei and Lactobacillus acidophilus, in the TNBS model of rat colitis." Journal of applied microbiology **103**(4): 836-844.

Perry, R. J., et al. (2016). "Acetate mediates a microbiome–brain–β-cell axis to promote metabolic syndrome." nature **534**(7606): 213-217.

Petrov, V., et al. (2017). "Comparison study of gut microbiota in case of Parkinson's disease and other neurological disorders." Bulletin of Siberian Medicine **15**(5): 113-125.

Portnoy, V. A., et al. (2011). "Adaptive laboratory evolution—harnessing the power of biology for metabolic engineering." Current opinion in biotechnology **22**(4): 590-594.

Pulikkan, J., et al. (2018). "Gut microbial dysbiosis in Indian children with autism spectrum disorders." Microbial ecology **76**(4): 1102-1114.

Rahimlou, M., et al. (2020). "Effects of long-term administration of Multi-Strain Probiotic on circulating levels of BDNF, NGF, IL-6 and mental health in patients with multiple sclerosis: a randomized, double-blind, placebo-controlled trial." Nutritional Neuroscience: 1-12.

Ren, L.-J., et al. (2013). "Impact of phosphate concentration on docosahexaenoic acid production and related enzyme activities in fermentation of Schizochytrium sp." Bioprocess and biosystems engineering **36**(9): 1177-1183.

Rezvani, F., et al. (2017). "Growth kinetic models of five species of Lactobacilli and lactose consumption in batch submerged culture." Brazilian Journal of Microbiology **48**: 251-258.

Rhoads, A., et al. (2015). "PacBio sequencing and its applications." Genomics, proteomics bioinformatics
13(5): 278-289.

Richter, M. and R. J. P. o. t. N. A. o. S. Rosselló-Móra (2009). "Shifting the genomic gold standard for the prokaryotic species definition." Proceedings of the National Academy of Sciences **106**(45): 19126-19131.

Riehle, M. M., et al. (2003). "Evolutionary changes in heat-inducible gene expression in lines of Escherichia coli adapted to high temperature." Physiological genomics **14**(1): 47-58.

Riley, M. A. J. A. r. o. g. (1998). "Molecular mechanisms of bacteriocin evolution." Annual review of genetics **32**(1): 255-278.

Rocha, E. P. and A. J. T. i. G. Danchin (2002). "Base composition bias might result from competition for metabolic resources." TRENDS in Genetics **18**(6): 291-294.

Rudolph, B., et al. (2010). "Evolution of Escherichia coli for growth at high temperatures." Journal of Biological Chemistry **285**(25): 19029-19034.

Rutherford, K., et al. (2000). "Artemis: sequence visualization and annotation." **16**(10): 944-945.

Salvetti, E., et al. (2012). "The genus Lactobacillus: a taxonomic update." Probiotics antimicrobial proteins **4**(4): 217-226.

Sanchez, S. and A. L. Demain (2002). "Metabolic regulation of fermentation processes." Enzyme and Microbial Technology **31**(7): 895-906.

Sanders, M. and T. J. J. o. d. s. Klaenhammer (2001). "Invited review: the scientific basis of Lactobacillus acidophilus NCFM functionality as a probiotic." Journal of Dairy Science **84**(2): 319-331.

Sanders, M. E. J. C. i. d. (2008). "Probiotics: definition, sources, selection, and uses." **46**(Supplement_2): S58-S61.

Sanger, F., et al. (1977). "DNA sequencing with chain-terminating inhibitors." Proceedings of the National Academy of Sciences **74**(12): 5463-5467.

Sanni, A., et al. (2002). "New efficient amylase-producing strains of Lactobacillus plantarum and L. fermentum isolated from different Nigerian traditional fermented foods." International journal of food microbiology **72**(1): 53-62.

Savolainen, H. and P. J. A. o. t. Pfäffli (1980). "Dose-dependent neurochemical changes during short-term inhalation exposure to m-xylene." Archives of toxicology **45**(2): 117-122.

Schillinger, U., et al. (2003). "Use of group-specific and RAPD-PCR analyses for rapid differentiation of Lactobacillus strains from probiotic yogurts." Current Microbiology **47**(6): 453-456.

Schwiertz, A., et al. (2010). "Microbiota and SCFA in lean and overweight healthy subjects." Obesity **18**(1): 190-195.

Seemann, T. J. B. (2014). "Prokka: rapid prokaryotic genome annotation." Bioinformatics **30**(14): 2068-2069.

Sender, R., et al. (2016). "Revised estimates for the number of human and bacteria cells in the body." PLoS biology **14**(8): e1002533.

Sengupta, R., et al. (2013). "The role of cell surface architecture of lactobacilli in host-microbe interactions in the gastrointestinal tract." **2013**.

Sequeira, L., et al. (1977). "Interaction of bacteria and host cell walls: its relation to mechanisms of induced resistance." **10**(1): 43-50.

Shapiro, J. A. (2009). "Revisiting the central dogma in the 21st century." Annals of the New York Academy of Sciences **1178**(1): 6-28.

Simon, O., et al. (2005). "Micro-organisms as feed additives-probiotics." Advances in pork Production **16**(2): 161.

Singhrao, S. K. and A. J. E. r. o. a.-i. t. Harding (2020). "Is Alzheimer's disease a polymicrobial host microbiome dysbiosis?" Expert review of anti-infective therapy **18**(4): 275-277.

Skoch, E., et al. (1981). "The effect of steam-conditioning rate on the pelleting process." Animal Feed Science Technology **6**(1): 83-90.

Soderlund, C., et al. (2011). "SyMAP v3. 4: a turnkey synteny system with application to plant genomes." Nucleic acids research: gkr123.

Steensels, J., et al. (2019). "Domestication of industrial microbes." Current biology **29**(10): R381-R393.

Sturme, M. H., et al. (2007). "Making sense of quorum sensing in lactobacilli: a special focus on Lactobacillus plantarum WCFS1." Microbiology **153**(12): 3939-3947.

Szajewska, H., et al. (2014). "Use of probiotics for management of acute gastroenteritis: a position paper by the ESPGHAN Working Group for Probiotics and Prebiotics." Journal of pediatric gastroenterology nutrition **58**(4): 531-539.

Tachedjian, G., et al. (2017). "The role of lactic acid production by probiotic Lactobacillus species in vaginal health." Research in microbiology **168**(9-10): 782-792.

Tahmourespour, A., et al. (2011). "Lactobacillus acidophilus-derived biosurfactant effect on gtfB and gtfC expression level in Streptococcus mutans biofilm cells." Brazilian Journal of Microbiology **42**(1): 330-339.

Thompson, J. D., et al. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic acids research **22**(22): 4673-4680.

Turnbaugh, P. J., et al. (2006). "An obesity-associated gut microbiome with increased capacity for energy harvest." nature **444**(7122): 1027-1031.

Ultee, E., et al. (2019). Stress-induced adaptive morphogenesis in bacteria. Advances in microbial physiology, Elsevier. **74:** 97-141.

van de Guchte, M., et al. (2006). "The complete genome sequence of Lactobacillus bulgaricus reveals extensive and ongoing reductive evolution." **103**(24): 9274-9279.

Van Erven, T. and P. J. I. T. o. I. T. Harremos (2014). "Rényi divergence and Kullback-Leibler divergence." IEEE Transactions on Information Theory **60**(7): 3797-3820.

van Heel, A. J., et al. (2018). "BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins." Nucleic acids research **46**(W1): W278-W281.

Vaser, R. and M. J. B. Šikić (2021). "Raven: a de novo genome assembler for long reads." BioRxiv: 2020.2008. 2007.242461.

Vaudagna, S. R., et al. (2002). "Sous vide cooked beef muscles: effects of low temperature–long time (LT–LT) treatments on their quality characteristics and storage stability." International Journal of Food Science Technology **37**(4): 425-441.

Ventura, M., et al. (2003). "The prophage sequences of Lactobacillus plantarum strain WCFS1." Virology **316**(2): 245-255.

Vernikos, G., et al. (2015). "Ten years of pan-genome analyses." Current opinion in microbiology **23**: 148-154.

Walker, B. J., et al. (2014). "Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement." PloS one **9**(11): e112963.

Wall, R., et al. (2012). "Contrasting effects of Bifidobacterium breve NCIMB 702258 and Bifidobacterium breve DPC 6330 on the composition of murine brain fatty acids and gut microbiota." The American journal of clinical nutrition **95**(5): 1278-1287.

Walls, A. B., et al. (2015). "The glutamine–glutamate/GABA cycle: function, regional differences in glutamate and GABA production and effects of interference with GABA metabolism." Neurochemical research **40**(2): 402-409.

Wang, J., et al. (2019). "CA-30, an oligosaccharide fraction derived from Liuwei Dihuang decoction, ameliorates cognitive deterioration via the intestinal microbiome in the senescence-accelerated mouse prone 8 strain." Aging **11**(11): 3463.

Wang, S., et al. (2020). "Lipoteichoic acid from the cell wall of a heat killed Lactobacillus paracasei D3-5 ameliorates aging-related leaky gut, inflammation and improves physical and cognitive functions: from C. elegans to mice." Geroscience **42**(1): 333-352.

Wick, R. R., et al. (2019). "Performance of neural network basecalling tools for Oxford Nanopore sequencing." Genome biology **20**(1): 1-10.

Wickham, H. J. W. I. R. C. S. (2011). "ggplot2." Wiley Interdisciplinary Reviews: Computational Statistics **3**(2): 180-185.

Wright, F. J. G. (1990). "The 'effective number of codons' used in a gene." Gene **87**(1): 23-29.

Wu, H., et al. (2012). "On the molecular mechanism of GC content variation among eubacterial genomes." Biology direct **7**(1): 1-16.

Xu, L., et al. (2019). "OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species." Nucleic acids research **47**(W1): W52-W58.

Xu, M., et al. (2016). "Encapsulation of Lactobacillus casei ATCC 393 cells and evaluation of their survival after freeze-drying, storage and under gastrointestinal conditions." Journal of food engineering **168**: 52-59.

Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Molecular biology and evolution **24**(8): 1586-1591.

Yoon, S.-H., et al. (2017). "A large-scale evaluation of algorithms to calculate average nucleotide identity." Antonie Van Leeuwenhoek **110**(10): 1281-1286.

Yunes, R., et al. (2016). "GABA production and structure of gadB/gadC genes in Lactobacillus and Bifidobacterium strains from human microbiota." Anaerobe **42**: 197-204.

Zhang, Y., et al. (2017). "Lactobacillus casei Zhang and vitamin K2 prevent intestinal tumorigenesis in mice via adiponectin-elevated different signaling pathways." Oncotarget **8**(15): 24719.

Zhang, Z., et al. (2019). "Demystifying the manipulation of host immunity, metabolism, and extraintestinal tumors by the gut microbiome." Signal transduction targeted therapy **4**(1): 1-34.

Zhao, X., et al. (2007). "GC content variability of eubacteria is governed by the pol III α subunit." Biochemical biophysical research communications **356**(1): 20-25.

Zhao, Y., et al. (2012). "PGAP: pan-genomes analysis pipeline." Bioinformatics **28**(3): 416-418.

Zheng, J., et al. (2020). "A taxonomic note on the genus Lactobacillus: Description of 23 novel genera, emended description of the genus Lactobacillus Beijerinck 1901, and union of Lactobacillaceae and Leuconostocaceae."

Zhou, X., et al. (2010). "The next-generation sequencing technology: a technology review and future perspective." Science China Life Sciences **53**(1): 44-57.

Zhuo, Q., et al. (2019). "Lysates of Lactobacillus acidophilus combined with CTLA-4-blocking antibodies enhance antitumor immunity in a mouse colon cancer model." Scientific reports **9**(1): 1-12.

국문초록

유산균의 진화적 및 기능적 특성에 대한 다중유전체학적 접근

전 수민

농생명공학부

서울대학교 대학원 농업생명과학대학

본 연구는 분자 수준에서 미생물의 진화와 그 기능성에 대한 포괄적 이해를 얻기 위해 실험적 증명 및 유전체 비교 분석을 수행되었다. 특히, 유산균은 인간에게 유용한 기능이 풍부한 종으로써 건강기능식품 등으로 많이 사용되고 있다. 유산균은 인간에게 유용한 기능성이 풍부하기 때문에, 학술적 및 상업적 연구 가치가 충분하다. 또한 유전체 크기가 작기 때문에 다른 개체에 비해 유전체 전체를 이해하기 용이하여, 유전체 연구에도 적합하다. 따라서 유산균의 유전체 연구는 인간에게 유용한 자원의 활용도를 높일 뿐만 아니라 복잡한 유전체를 가진 고등생물의 유전적 이해를 하는 데에도 기여를 할 것이다. 따라서

본 연구는 다양한 유전체 분석을 통해 인간에게 유용한 유산균의 진화와 기능에 대한 다면적 이해를 제공하고자 수행되었다.

2 장에서는 기능성 유산균인 *Lactiplantibacillus plantarum* GB-LP3 의 전장 유전체 서열을 해독하고, 기존에 해독된 *L. plantarum* 전장 유전체들과 비교하여 진화적 특성을 판별하였다. Infant fecal samples 에서 동정된 ZJ316 의 유전체와 가장 가까운 진화적 거리를 가지고 있으며, 다수의 기능성 유전자 및 진화적으로 가속화된 ATP transporter 를 보유함을 확인하였다. 이를 토대로 발효식품이라는 특수한 환경 속에서 *L. plantarum* 의 적응 방식을 추론할 수 있었다.

3 장에서는 건강기능식품으로 널리 사용 중인 *Lactobacillus bulgaricus* 와 *Limosilactobacillus fermentum* 이 다른 유산균 종들에 비해 Genome size 대비 높은 GC content 를 보유함을 확인하였다. 높은 GC content 는 아미노산을 암호화하는 triplet code 의 세 번째 nucleotide 의 차이 때문임을 확인하고 이로 인해 발생하는 에너지의 차이를 비교하였다. 이를 통해 *L. bulgaricus* 와 *L. fermentum* 가 환경에 적응하기 위해 높은 GC content 를 갖는 쪽으로 진화하였음을 추론하였다.

4 장에서는 진화압에 대한 표현형과 유전자형의 변화를 확인하기 위해, 인위적으로 고온에 노출시켜 내열성을 높인

Lactobacillus acidophilus 균주를 개발하였다. 열적응 균주는 야생형에 비해 65 도 이상의 고온에서 생존율이 유의하게 증가하였다. 전장 유전체 비교를 통해 세포벽에 관련된 유전자 부근에서 2 개의 SNP 를 확인하였다. 이를 토대로 *L. acidophilus* 균주가 열자극에 적응하기 위하여 세포벽을 단단하게 하는 방향으로 진화하였음을 제시하였다.

5 장에서는 *Lactobacillus acidophilus*, *Lacticaseibacillus paracasei*, *Lacticaseibacillus rhamnosus* 를 8 주간 투여한 쥐의 인지능력을 평가하고 장내미생물 조성의 변화를 비교하여 LAB 가 *in vivo* 실험동물에 미치는 영향을 확인하였다. 실험군 중 *L. acidophilus* 를 먹인 균주에서 가장 높은 인지능력의 향상을 보였으며 이와 함께 두 그룹간 장내미생물 균총 비교에서 16 개의 박테리아 종이 유의미하게 차이를 나타냈다. 비율이 변한 박테리아의 상당수가 동물의 뇌에 작용하는 신경물질 합성에 필요한 물질 생산에 관여하는 것으로, 장내에 늘어난 *L. acidophilus* 균총 증가의 영향으로 신경물질의 합성량이 증가했고 그로 인해 인지능력이 향상되었음을 추론 및 제시하였다.

본 논문의 2 장부터 5 장까지는 3 세대 염기서열분석과 유전체 분석을 통해 유산균의 진화와 기능적 특성을 제시하였다. 구체적으로 전장 유전체 해독, 계통수 작성, 유전체 비교 분석,

metagenome 분석을 수행하여 유산균에 대한 이해에 적용하였다. 이러한 연구를 통해 유전자 분석을 통해 유산균의 기능적, 진화적 특성을 제시할 수 있을 뿐만 아니라 실험을 통해 기대되는 기능을 규명하고 유전적 요인을 유추할 수 있었다. 이 연구를 통해 미생물의 특성과 유전체 분석에 대한 포괄적인 이해가 깊어지기를 바란다.

주요어

유전체 해독, 3 세대 염기서열 해독, 미생물 진화, 유산균, 락토바실러스