J. Microbiol. Biotechnol. (2018), 28(8), 1352–1359
https://doi.org/10.4014/jmb.1803.03029

jmb

Pan-Genomics of *Lactobacillus plantarum* Revealed Group-Specific Genomic Profiles without Habitat Association S

Sukjung Choi[1], Gwi-Deuk Jin[1,2], Jongbin Park[1,2], Inhwan You[1,2], and Eun Bae Kim[1,2,3*]

[1]*Laboratory of Microbial Genomics and Big Data, College of Animal Life Sciences, Kangwon National University, Chuncheon 24341, Republic of Korea*
[2]*Division of Applied Animal Science, College of Animal Life Sciences, Kangwon National University, Chuncheon 24341, Republic of Korea*
[3]*Department of Animal Life Science, College of Animal Life Sciences, Kangwon National University, Chuncheon 24341, Republic of Korea*

Received: March 21, 2018
Revised: May 24, 2018
Accepted: May 25, 2018

First published online
June 8, 2018

**Corresponding author*
Phone: +82-33-250-8642;
Fax: +82-33-259-5574;
E-mail: itanimal@kangwon.ac.kr

S upplementary data for this paper are available on-line only at http://jmb.or.kr.

pISSN 1017-7825, eISSN 1738-8872

Lactobacillus plantarum is a lactic acid bacterium that promotes animal intestinal health as a probiotic and is found in a wide variety of habitats. Here, we investigated the genomic features of different clusters of *L. plantarum* strains via pan-genomic analysis. We compared the genomes of 108 *L. plantarum* strains that were available from the NCBI GenBank database. These genomes were 2.9−3.7 Mbp in size and 44−45% in G+C content. A total of 8,847 orthologs were collected, and 1,709 genes were identified to be shared as core genes by all the strains analyzed. On the basis of SNPs from the core genes, 108 strains were clustered into five major groups (G1–G5) that are different from previous reports and are not clearly associated with habitats. Analysis of group-specific enriched or depleted genes revealed that G1 and G2 were rich in genes for carbohydrate utilization (L-arabinose, L-rhamnose, and fructo-oligosaccharides) and that G3, G4, and G5 possessed more genes for the restriction-modification system and MazEF toxin-antitoxin. These results indicate that there are critical differences in gene content and survival strategies among genetically clustered *L. plantarum* strains, regardless of habitats.

Keywords: *Lactobacillus plantarum*, SNP, animal, plant, comparative genomics, pan-genome

Introduction

Lactobacillus plantarum, a gram-positive and acid-tolerant bacterium, is an economically interesting member of the lactic acid bacteria family [1, 2]. This is because *L. plantarum* strains have been found in a wider variety of habitats compared with other lactobacilli. Representative habitats of *L. plantarum* include the gastrointestinal tracts (GITs) of insects and animals, as well as various fermented food products such as dairy products, fermented beverages, meat products, pickles, and kimchi [1, 2]. Since *L. plantarum* can survive in the GIT and has excellent long-term fixability, they have become increasingly significant in maintaining intestinal health [3, 4]. Indeed, certain *L. plantarum* strains have been commercialized as probiotics owing to their health benefits [5, 6].

There have been several genotypic and phenotypic comparative studies on *L. plantarum* strains over the past 10 years. Molenaar *et al.* [7] investigated the gene categories of 20 *L. plantarum* strains. They found that genes involved in the synthesis or degradation of proteins and lipids were largely conserved, but the genes involved in sugar transport and catabolism were highly variable between strains. In another study, 24 phenotypes in 185 *L. plantarum* strains were evaluated for their fermentation and growth characteristics [8]. However, the genotypic associations with such phenotypes were not fully considered and investigated. A recent phylogenetic analysis of the core genome revealed the absence of habitat-related phylogenetic groups [9].

Despite such comparative studies, much about *L. plantarum* remains unclear. Firstly, the number of genomes that was analyzed was not sufficient. Recently, pan-genomic analyses of *Enterococcus faecium* substantially increased the number of strains studied, thereby improving the general understanding about the organism and providing new insights [10−13]. As the number of *L. plantarum* genomes has increased in the National Center for Biotechnology Information (NCBI) GenBank database, it has become necessary to subdivide the population into new clusters and obtain information based on the new divisions. Secondly, the previous phylogenetic analysis was mainly based on amino acid sequences of the core genes or on the absence or presence of genes, and thus failed to account for single nucleotide polymorphisms (SNPs), which are known to increase the resolution in phylogeny. Finally, it was previously described that *L. plantarum* has a nomadic lifestyle [9]; however, this conclusion was drawn using only 54 strains and the study focused only on habitat associations without considering the genomic features of each cluster.

In this study, we compared the genomes of 108 *L. plantarum* strains for phylogenetic clustering based on SNPs and focused on the genomic features of phylogenetic clusters, rather than habitat-associated groups, to determine whether each phylogenetic cluster possesses cluster-specific genomic features that may be independent of habitats.

Materials and Methods

Genomes and Strains Used in This Study

The genomes of 108 *L. plantarum* strains used in this study were obtained from the NCBI database. Complete and incomplete genomes were retrieved in August 2016 (Table S1). Habitat classification information is added as a footnote to the table.

Collection of *L. plantarum* Orthologous CDS

Genomes were re-annotated by using the RAST server [14] with default options for bacteria. To obtain orthologs, we extracted protein-coding sequences (CDSs) from re-annotated GenBank files. Such CDSs were mutually aligned using GASSST [15] with parameters of ≥95% sequence identity and a sensitivity level of 5 (maximum). Each collection of clustered CDS was assembled to make a consensus orthologous CDS.

Defining Gene Presence and Absence

The contigs for each genome were fragmented into 50-bp sequences by 7-bp intervals. Each fragment was aligned onto *L. plantarum* orthologous CDSs by using GASSST with parameters of ≥95% sequence identity and a sensitivity level of 5 (maximum). If a gene was covered by fragments greater than 90% of the CDS length, it was recognized and recorded to be present in a particular genome. In this way, all orthologous genes were evaluated for their presence/absence in each genome. Genes that are commonly shared by all genomes are regarded as core genes. This process allowed for the identification of a number of genes in each phylogenetic group. The significance of different gene frequencies by group was evaluated by Fisher's exact test. Nonsignificant genes were omitted ($p > 0.05$) and the odds ratios were calculated to determine group-specific enriched or depleted genes (Table S2).

Phylogenetic Analysis

To analyze phylogenetic similarities among the 108 *L. plantarum* strains, we used SNP data found in the core genes. We defined 1,709 genes that are commonly found in all genomes as core genes. Among the 1,709 core genes, only 1,430 genes were used for SNP analysis because of incomplete genome assemblies, and the other 279 genes were excluded. To collect core gene sequences from each genome, the ortholog sequences were aligned to each genome using GASSST under the 90% sequence similarity option. The aligned regions in each genome were used for SNP selection. The collected core gene sequences from each genome were aligned among the 108 strains using the alignment tool MUSCLE [16]. On the basis of the core gene alignment, we were able to detect SNPs. Polymorphic sites were detected and collected for construction of a phylogenetic tree using MEGA7 [17]. Evolutionary trajectories were inferred by the neighbor-joining method [18] based on 1,000 bootstrap replicates [19]. The evolutionary distances between strains were computed using the maximum composite likelihood method [20] and were used to infer phylogenetic trees.

Comparison of Gene Composition

After re-annotation by the RAST server, we downloaded a gene functional information file called subsystems provided by the RAST server. The number of genes for each subsystem category was counted using in-house Perl scripts. We generated one table by counting the number of subsystem genes for each strain. Student's *t*-test was performed to determine the differences in the number of genes between the groups.

Results

Pan-Genomic Statistics of 108 *L. plantarum* Strains

We analyzed the genomes of 108 *L. plantarum* strains obtained from the NCBI GenBank database. Strains of animal ($n = 57$), plant ($n = 38$), gut ($n = 27$, a subset of animal strains), dairy product and breast (shortly dairy, $n = 16$, a subset of animal strains), meat product ($n = 7$, a subset of animal strains), and unknown ($n = 13$) origins were included (Tables S1 and S3). We identified 8,847 orthologous CDSs (Table S4) and found that 1,709 core genes (Tables S4 and S5) were shared by all strains. The number of core genes was not critically biased to any specific strain when we subsampled different numbers of strains (Fig. S1). The average genome size and G+C content of the 108 *L. plantarum* strains were 3.3 ± 0.1 Mbp and 44.4 ± 0.2%,

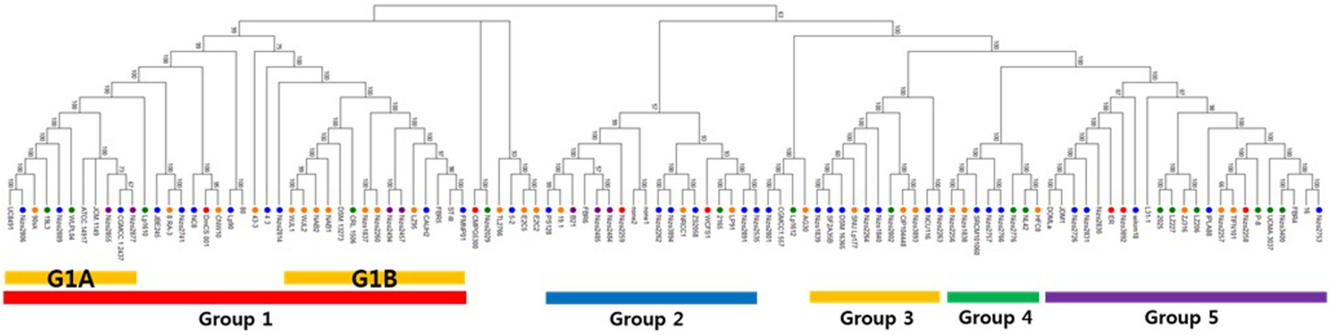


Fig. 1. Evolutionary relationships of 108 *Lactobacillus plantarum* strains based on 85,270 SNPs from 1,430 core genes.
The evolutionary history was inferred using the neighbor-joining method. The bootstrap consensus tree inferred from 1,000 replicates was taken to represent the evolutionary history of the taxa analyzed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1,000 replicates) are shown next to the branches. Evolutionary analyses were conducted using MEGA7. Each strain origin is displayed by circles of different colors. ● Animal-derived strains; ● plant; ● dairy; ● gut; ● meat.

respectively. The number of CDSs per genome ranged from 2,711 to 3,595 (3,103 ± 163).

Genomic Statistics of SNP-Based Phylogenetic Groups

Unlike previous studies, we accounted for SNPs during the phylogenetic tree construction to provide more comprehensive and informative phylogenetic relationships among the 108 strains. A total of 85,270 SNPs were identified from the 1,430 core genes. On average, SNPs were found 67 times per kilobase pair per gene (66.6 ± 31.7, Table S5). The SNPs were further used for the construction of a phylogenetic tree (Fig. 1). Five major groups (G1–G5) and two subgroups within G1 were identified. As the G1A and G1B subgroups had similar genomic contents when their RAST annotation profiles were compared (data not shown), they were merged into the major group G1. Genomic parameters (genome size, number of CDSs, and G+C content) among the five groups were compared (Table 1). A significant difference was found in the G+C content of accessory genes in all groups ($p < 0.001$). No statistical differences were found in the genome size and in the number of CDSs among the five groups.

Habitat Association with *L. plantarum* Groups

In a recent study [9], no association between habitats and phylogenetic groups of *L. plantarum* was observed. However, specific statistics were not reported to support this. As we observed new phylogenetic groups by employing a different method based on SNPs in this study, we examined the association of the clusters with isolation origins (Table S6). Strains from meat products and plants were slightly enriched in G2 at $p = 0.19$ and in G3 at $p = 0.27$, respectively. Strains from meat products were not found in the G3, G4, and G5 groups. We were not able to find any clear statistical evidence for environmental association with new phylogenetic clusters. We also analyzed genomic parameters such as genome size, number of CDSs, and G+C content (%) among the five origins (Table S3) and

Table 1. Genomic comparisons by each group.

Groups	No. of strains	Genome size (kbp)	No. of CDSs	G+C content(%) in core genes	G+C content(%) in accessory genes
G1	38	3,299±92	3,118±163	46.0±3.2	41.6±5.3^{c}
G2	18	3,299±165	3,108±216	46.0±3.2	42.0±5.3^{b}
G3	11	3,246±85	3,066±121	46.0±3.2	42.7±5.4^{a}
G4	8	3,187±144	3,053±126	46.0±3.2	42.9±5.0^{a}
G5	23	3,250±97	3,094±135	46.0±3.2	42.2±5.2^{b}
p Values	Not applicable	0.0684	0.788	Not applicable	***

*Data are presented as the mean ± standard deviation. *p* values were calculated by ANOVA test.

$^{a-c}$Values of the G+C content in accessory genes with different superscripts are significantly different ($p < 0.05$). ***$p < 0.001$.

found no statistically significant differences. In addition, we tried to identify any genes that were differentially enriched or depleted in any of the origins at $p < 0.05$ (Table S7). A total of 77 and 13 genes were identified to be habitat-enriched (21, 7, and 40 genes from gut, dairy, and meat origins, respectively) and habitat-depleted (4 and 6 genes from plant and gut origins, respectively) genes, respectively. Owing to the limited number of meat-derived strains, most of the enriched genes were from the meat origin. Any animal-specific enriched or depleted genes were not found. We also found phage-associated genes, 19 (5 from gut and 2 from dairy origins) of which were enriched and two were depleted. Among the 13 depleted genes, four fructose-related PTS genes were identified and they were all of gut origin.

Group-Specific Enriched or Depleted Gene Categories

As habitat-association was not clear, we focused on the characterization of the five newly clustered major phylogenetic groups (G1–G5). The RAST Annotation Server supports a classification for gene categories called subsystems. The subsystems were compared among the five groups. Genes belonging to 28 subsystems were differentially enriched or depleted (Table 2, Fig. 2). Both G1 and G2 were enriched for the genes associated with carbohydrate utilization when compared with the expected average for all strains. However, G3, G4, and G5 had less genes than expected. G1 was the only group that had genes related to xylose utilization. G2 had the highest number of genes for inositol catabolism ($p < 0.05$), whereas G5 had the least number of genes for carbohydrate utilization.

On the other hand, both G1 and G2 were poor in genes related to the MazEF toxin-antitoxin and toxin-antitoxin replicon stabilization systems. Such genes were relatively more abundant in G3 and G5. In addition, the number of genes for restriction-modification systems, as well as cadmium, mercury, and arsenic resistance systems, were significantly elevated in G3 and G5. G2 was the most abundant in the two nitrogen metabolism subsystems, but was poor in the two sulfur metabolism subsystems. The results from the individual gene analysis agreed with these observations (Fig. 3). Six genes associated with carbohydrate utilization were enriched in G1. The restriction-modification system, which is a bacterial defense mechanism, and MazF, which prompts cell death, were found predominantly in G5. G1 and G5 were inverse to each other in terms of

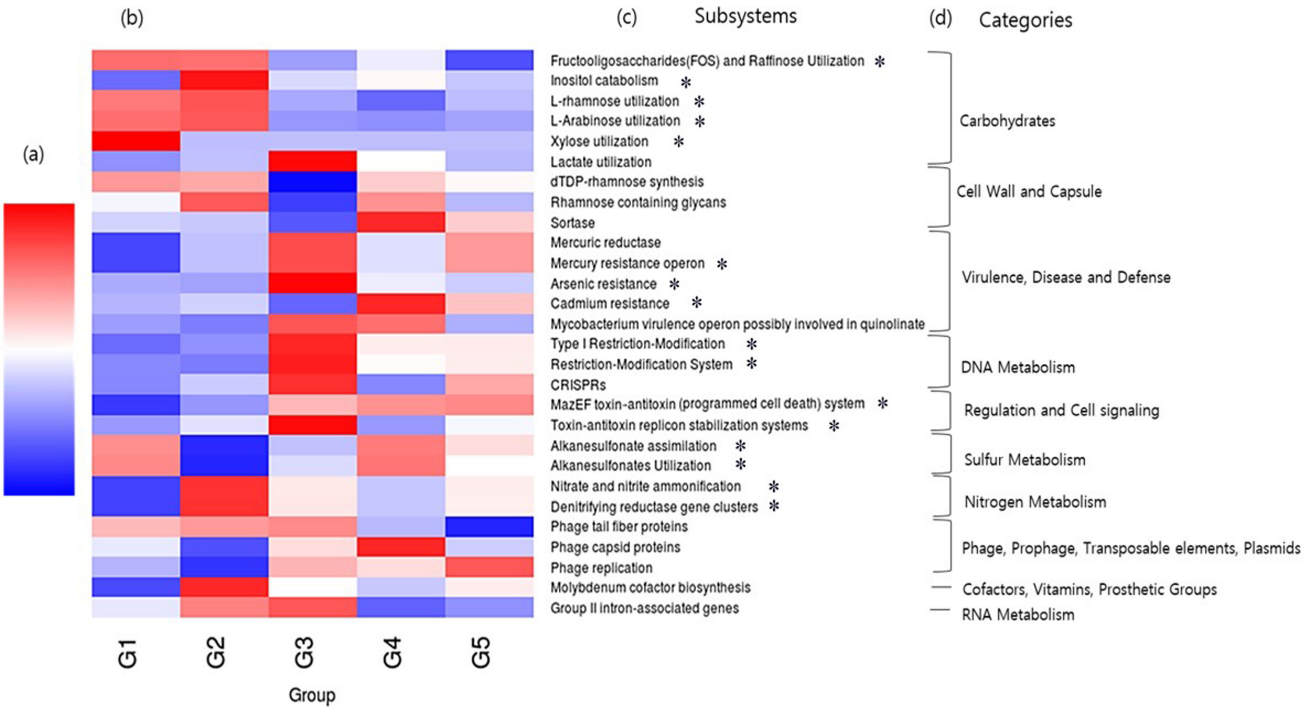


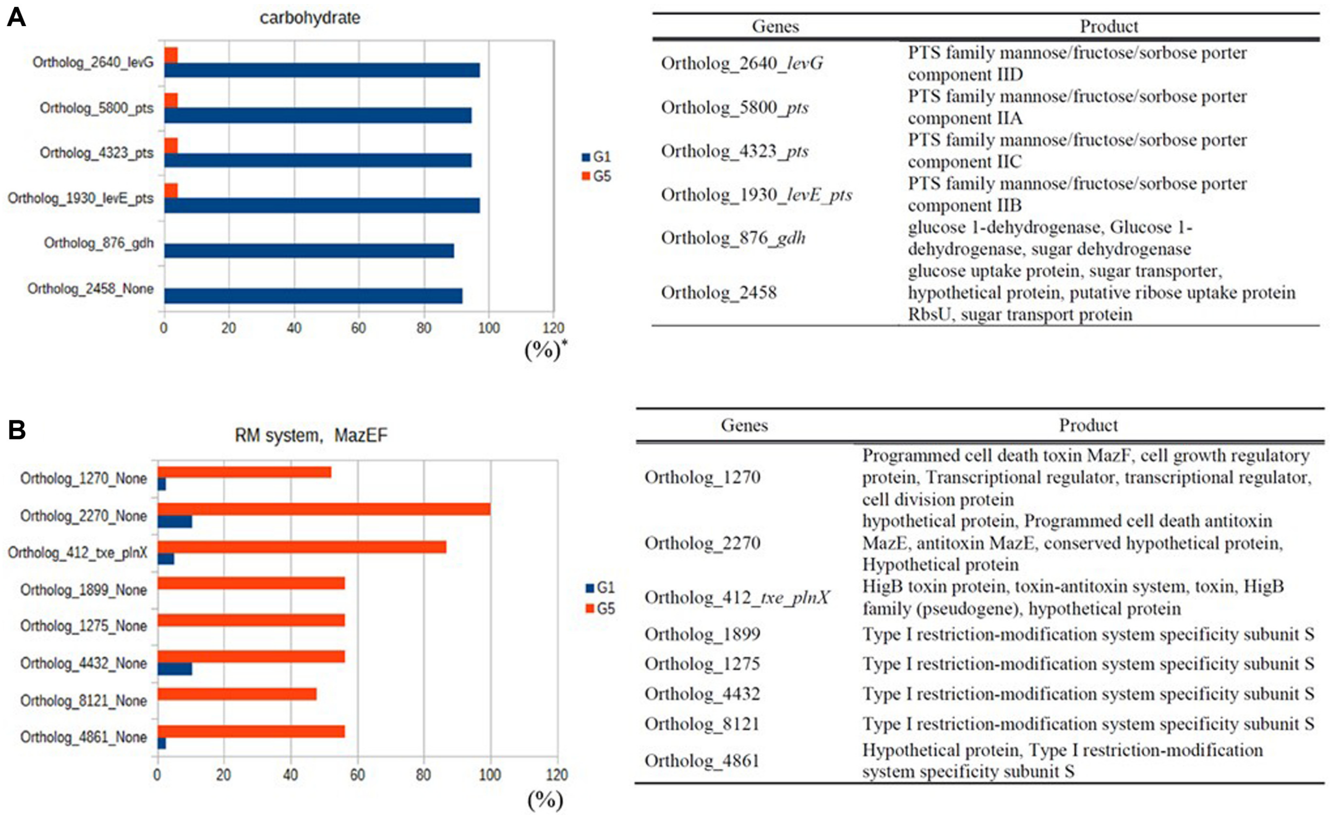
Fig. 2. Heatmap for gene abundance of functional categories among *L. plantarum* groups.

The number of genes in each cell was scaled on the horizontal axis to allow for group comparisons. Red indicates groups with more genes than average values in each row, whereas blue indicates those with less. An asterisk (*) is displayed on subsystems that show significant differences by group.

Table 2. Differences in gene categories (subsystems) among five *L. plantarum* groups.

Categories	Subsystems	All	G1		G2		G3		G4		G5	
		Mean	Mean	*p*	Mean	*p*	Mean	*p*	Mean	*p*	Mean	*p*
Carbohydrates	Fructooligosaccharides and raffinose utilization	8.6 ± 6.2	10.8 ± 5.3	*	10.8 ± 5.1	NS	6.3 ± 6.3	NS	7.8 ± 5.2	NS	4.8 ± 6.2	*
	Inositol catabolism	1.5 ± 3.0	0.2 ± 1.1	***	4.1 ± 3.8	*	1.3 ± 2.8	NS	1.8 ± 3.2	NS	1.1 ± 3.0	NS
	L-Rhamnose utilization	7.8 ± 5.8	10.7 ± 3.8	***	11.9 ± 1.4	***	3.6 ± 6.1	*	1.5 ± 4.2	**	4.2 ± 5.9	*
	L-Arabinose utilization	3.4 ± 2.4	4.7 ± 1.4	***	5.0 ± 0.0	***	1.4 ± 2.3	*	1.3 ± 2.3	*	1.5 ± 2.4	**
	Xylose utilization	0.1 ± 1.0	0.3 ± 1.6	NS	0.0 ± 0.0	NS	0.0 ± 0.0	NS	0.0 ± 0.0	NS	0.0 ± 0.0	NS
	Lactate utilization	0.1 ± 0.5	0.0 ± 0.0	**	0.1 ± 0.5	NS	0.8 ± 1.2	NS	0.3 ± 0.7	NS	0.1 ± 0.4	NS
Cell wall and capsule	dTDP-rhamnose synthesis	3.0 ± 3.3	3.2 ± 4.0	NS	3.1 ± 2.6	NS	2.1 ± 2.9	NS	3.0 ± 2.5	NS	2.9 ± 3.1	NS
	Rhamnose-containing glycans	5.8 ± 6.0	5.6 ± 6.6	NS	7.1 ± 6.0	NS	3.9 ± 5.4	NS	6.6 ± 5.5	NS	5.0 ± 5.4	NS
	Sortase	1.3 ± 0.8	1.2 ± 0.6	NS	1.2 ± 0.5	NS	1.0 ± 0.0	***	1.8 ± 0.5	*	1.4 ± 1.3	NS
Virulence, disease, and defense	Mercuric reductase	0.1 ± 0.4	0.0 ± 0.2	*	0.1 ± 0.5	NS	0.3 ± 0.5	NS	0.1 ± 0.4	NS	0.2 ± 0.4	NS
	Mercury resistance operon	0.1 ± 0.4	0.0 ± 0.2	*	0.1 ± 0.5	NS	0.3 ± 0.5	NS	0.1 ± 0.4	NS	0.2 ± 0.4	NS
	Arsenic resistance	0.9 ± 1.9	0.6 ± 1.2	NS	0.5 ± 1.5	NS	2.8 ± 3.5	NS	1.0 ± 1.4	NS	0.8 ± 1.3	NS
	Cadmium resistance	0.1 ± 0.4	0.1 ± 0.3	NS	0.1 ± 0.3	NS	0.0 ± 0.0	***	0.4 ± 0.5	NS	0.2 ± 0.5	NS
	Mycobacterium virulence operon possibly involved in quinolinate biosynthesis	0.2 ± 0.7	0.1 ± 0.5	NS	0.0 ± 0.0	**	0.8 ± 1.4	NS	0.8 ± 1.4	NS	0.1 ± 0.6	NS
DNA metabolism	Type I restriction-modification	2.9 ± 2.9	2.3 ± 2.7	NS	2.5 ± 2.8	NS	4.7 ± 3.1	NS	3.4 ± 3.1	NS	3.4 ± 3.4	NS
	Restriction-modification system	3.1 ± 3.0	2.6 ± 2.6	NS	2.5 ± 2.8	NS	4.8 ± 3.1	NS	3.4 ± 3.1	NS	3.5 ± 3.3	NS
	CRISPRs	0.4 ± 1.1	0.0 ± 0.0	***	0.2 ± 0.7	NS	0.8 ± 2.1	NS	0.0 ± 0.0	***	0.5 ± 1.2	NS
Regulation and cell signaling	MazEF toxin-antitoxin (programmed cell death) system	1.2 ± 1.7	0.1 ± 0.5	***	0.8 ± 1.5	NS	2.1 ± 1.7	NS	2.4 ± 1.5	NS	2.4 ± 2.0	*
	Toxin-antitoxin replicon stabilization system	0.2 ± 0.7	0.0 ± 0.0	**	0.2 ± 0.7	NS	0.8 ± 1.2	NS	0.0 ± 0.0	**	0.2 ± 0.7	NS
Sulfur metabolism	Alkanesulfonate assimilation	2.6 ± 1.8	3.6 ± 0.5	***	0.0 ± 0.0	***	1.6 ± 1.6	NS	3.8 ± 0.7	**	2.7 ± 2.0	NS
	Alkanesulfonates utilization	1.4 ± 0.9	2.0 ± 0.2	***	0.0 ± 0.0	***	1.1 ± 1.0	NS	2.1 ± 0.4	***	1.3 ± 1.0	NS
Nitrogen metabolism	Nitrate and nitrite ammonification	1.5 ± 2.4	0.0 ± 0.0	***	3.7 ± 2.4	**	2.0 ± 2.8	NS	1.3 ± 2.3	NS	2.0 ± 2.5	NS
	Denitrifying reductase gene clusters	1.2 ± 1.9	0.0 ± 0.0	***	3.0 ± 2.0	**	1.6 ± 2.3	NS	1.0 ± 1.9	NS	1.6 ± 2.0	NS
Phages, prophages, transposable elements, plasmids	Phage tail fiber proteins	1.1 ± 1.1	1.4 ± 1.3	NS	1.5 ± 1.1	NS	1.6 ± 1.0	NS	0.9 ± 1.1	NS	0.4 ± 0.6	***
	Phage capsid proteins	4.5 ± 3.1	4.3 ± 3.0	NS	2.8 ± 1.4	***	4.9 ± 2.0	NS	6.8 ± 3.8	NS	4.0 ± 2.8	NS
	Phage replication	2.4 ± 2.1	2.1 ± 1.9	NS	1.5 ± 1.9	NS	2.8 ± 2.0	NS	2.6 ± 1.7	NS	3.3 ± 2.1	NS
Cofactors, vitamins, prosthetic groups, pigments	Molybdenum cofactor biosynthesis	5.7 ± 4.2	3.1 ± 0.3	***	9.9 ± 4.4	***	6.3 ± 4.5	NS	5.3 ± 4.2	NS	6.5 ± 4.5	NS
RNA metabolism	Group II intron-associated genes	0.4 ± 0.6	0.3 ± 0.5	NS	0.6 ± 0.8	NS	0.6 ± 0.7	NS	0.0 ± 0.0	***	0.1 ± 0.3	**

Comparisons between a specific group and the average of all strains were examined, and statistical significance was denoted by ***, **, and * for $p < 0.001$, $p < 0.01$, and $p < 0.05$, respectively. NS indicates no significance.



Genes	Product
Ortholog_2640_*levG*	PTS family mannose/fructose/sorbose porter component IID
Ortholog_5800_*pts*	PTS family mannose/fructose/sorbose porter component IIA
Ortholog_4323_*pts*	PTS family mannose/fructose/sorbose porter component IIC
Ortholog_1930_*levE_pts*	PTS family mannose/fructose/sorbose porter component IIB
Ortholog_876_*gdh*	glucose 1-dehydrogenase, Glucose 1-dehydrogenase, sugar dehydrogenase
Ortholog_2458	glucose uptake protein, sugar transporter, hypothetical protein, putative ribose uptake protein RbsU, sugar transport protein

Genes	Product
Ortholog_1270	Programmed cell death toxin MazF, cell growth regulatory protein, Transcriptional regulator, transcriptional regulator, cell division protein
Ortholog_2270	hypothetical protein, Programmed cell death antitoxin MazE, antitoxin MazE, conserved hypothetical protein, Hypothetical protein
Ortholog_412_*txe_plnX*	HigB toxin protein, toxin-antitoxin system, toxin, HigB family (pseudogene), hypothetical protein
Ortholog_1899	Type I restriction-modification system specificity subunit S
Ortholog_1275	Type I restriction-modification system specificity subunit S
Ortholog_4432	Type I restriction-modification system specificity subunit S
Ortholog_8121	Type I restriction-modification system specificity subunit S
Ortholog_4861	Hypothetical protein, Type I restriction-modification system specificity subunit S

Fig. 3. Enriched and depleted genes in the G1 and G5 groups.

(**A**) Six genes associated with carbohydrate metabolism were enriched in G1. (**B**) Eight genes related to the MazEF toxin-antitoxin system and restriction-modification system were enriched in G5. * The probability that the gene exists in G1 and G5.

enrichment of genes for carbohydrate utilization and cell self-defense and death, respectively.

Group-Specific Enriched or Depleted Genes and Gene Ontology

Group-specific genes (Tables 3 and S2) were further analyzed by their gene ontology categories (Table S9). All the group-specific genes were accessory genes. Many genes were enriched in one group and depleted in another. We identified 1,534 genes that were either enriched or depleted in one of the five groups (Table 3). Unfortunately, only a limited number of genes were classified by gene ontology. G3 had the highest number of enriched genes, whereas G5 had the lowest number of depleted genes. Gene ontology analysis showed that certain gene categories were enriched (Table S9). G1-specific genes were enriched in categories such as transport of organic or amino acids and chitinase activity. Categories such as nitrate reductase activity and carbohydrate transmembrane transporter activity were depleted in G1-specific genes, but were enriched in G2-specific genes. Twelve G3-specific genes were also found in the kinase activity category.

Discussion

In this study, we analyzed *L. plantarum* genomes to show major differences among five newly defined phylogenetic clusters. Each group had distinct genomic contents,

Table 3. Number of group-specific enriched or depleted genes.

Origins	G1	G2	G3	G4	G5
Enriched genes					
Total	141	105	621	215	104
Gene with GO ID	18	27	114	28	7
Depleted genes					
Total	212	36	85	56	253
Gene with GO ID	27	1	10	11	27

Significance was examined at $p < 0.05$ by Fisher's exact test when compared with the gene frequencies in all strains. Gene ontology (GO) ID was assigned by the RAST annotation server. Refer to Table S2 and Table S9 for more details.

including group-specific genes and gene categories. A recent study showed a lack of habitat association, reflecting the nomadic lifestyle of *L. plantarum* [9]. However, we identified that certain origins were weakly associated with two SNP-based groups (meat with G2 and plant with G3, Table S6). Moreover, each origin had origin-specific genes and gene categories.

The SNP ratios of genes associated with either cell membrane component or transport activity were high, whereas the SNPs associated with either ribosomes or RNA polymerases were low (Table S5). We found that there was no correlation between SNP frequency and gene length ($R = 0.14$, $p < 0.001$). Ribosomes are essential in living organisms, and it has been well established that genes for ribosomal proteins are conserved in bacterial genera. In contrast, cell membrane components or transporters are important for adaptation to the surrounding environments [21]. Thus, we think that our phylogenetic tree reflects variations in environmental factors rather than other factors shared among *L. plantarum* strains.

For the five phylogenetic clusters, we were unable to find any critical differences in genome size, number of CDSs, and G+C content (Table 1). However, we found that genome contents were different among the five groups. The strains in G1 and G2 appeared to have opposite tendencies to strains in G4 and G5. G1 and G2 had a higher capacity to metabolize various carbohydrates such as glucose, fructose, galactose, and lactose, but the genes related to these processes were rare in G4 and G5 (Table 2 and Fig. 2). In contrast, the MazEF toxin-antitoxin system, which is involved in apoptosis, was deficient in G1 and G2 but was identified significantly more frequently in G5 than in the other groups (Odds Ratio > 4.0). MazE is antitoxic and MazF is toxic, and these two are co-expressed and interact with each other. The activity of MazF is neutralized by the antitoxic effects of MazE. [22, 23]. In starved and stressed bacteria, suicide mechanisms may be triggered at high cell densities in order to enable the use of dead cells as alternative and emergency sources of nutrients [24]. Another notable feature of strains in G4 and G5 is that they were rich in genes related to type 1 restriction-modification systems (Odds Ratio > 10 and Odds Ratio > 3.5, respectively), which are systems that allow bacteria to distinguish and destroy foreign DNA entering the cell, such as those from bacteriophages. There are two enzymes involved in this process: a restriction endonuclease that cleaves foreign DNA, and a modification methyltransferase that protects the host DNA [25, 26]. Although most *L. plantarum* strains are capable of growth using a wide variety of carbohydrate sources [27], strains in G4 and G5 unusually lacked genes for carbohydrate transport and degradation. It can be problematic for microorganisms in these groups to obtain sufficient quantities of nutrients, which thus results in cell mortality; however, this allows surviving cells to use dead cells as metabolite sources [24]. In addition, these groups may also be able to create a defense mechanism to protect themselves by distinguishing endogenous DNA from foreign DNA [25, 26]. Such protective strains may lose chances to obtain foreign genes that may be beneficial to the host. To test this hypothesis, more studies are required.

From our pan-genomic analysis, we identified that certain genes and ontology categories were either group-specifically or origin-specifically enriched or depleted. However, we were unable to fully understand why such variations happened and what was beneficial to each strain. For a better understanding, we need to further validate the association between such categories/genes and phenotypes/origins. Such efforts will be helpful to obtain genetic makers for better probiotic or commensal *L. plantarum* strains.

Acknowledgments

This study was supported by a grant from the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2016R1C1B2016246). Gwi-Deuk Jin, Jongbin Park, and Inhwan You were supported by the Brain Korea 21 Plus Program.

Conflict of Interest

The authors have no financial conflicts of interest to declare.

References

1. Holzapfel W, Wood BJ. 2012. *The Genera of Lactic Acid Bacteria.* Springer Science & Business Media, Berlin.
2. Makarova KS, Koonin EV. 2007. Evolutionary genomics of lactic acid bacteria. *J. Bacteriol.* **189:** 1199-1208.
3. De Vries MC, Vaughan EE, Kleerebezem M, de Vos WM. 2006. *Lactobacillus plantarum* – survival, functional and potential probiotic properties in the human intestinal tract. *Int. Dairy J.* **16:** 1018-1028.
4. Zago M, Fornasari ME, Carminati D, Burns P, Suàrez V, Vinderola G, *et al.* 2011. Characterization and probiotic potential of *Lactobacillus plantarum* strains isolated from cheeses. *Food Microbiol.* **28:** 1033-1040.

5. Saxelin M, Tynkkynen S, Mattila-Sandholm T, de Vos WM. 2005. Probiotic and other functional microbes: from markets to mechanisms. *Curr. Opin. Biotechnol.* **16:** 204-211.
6. Seddik HA, Bendali F, Gancel F, Fliss I, Spano G, Drider D. 2017. *Lactobacillus plantarum* and its probiotic and food potentialities. *Probiotics Antimicrob. Proteins* **9:** 111-122.
7. Molenaar D, Bringel F, Schuren FH, de Vos WM, Siezen RJ, Kleerebezem M. 2005. Exploring *Lactobacillus plantarum* genome diversity by using microarrays. *J. Bacteriol.* **187:** 6119-6127.
8. Siezen RJ, Tzeneva VA, Castioni A, Wels M, Phan HT, Rademaker JL, *et al*. 2010. Phenotypic and genomic diversity of *Lactobacillus plantarum* strains isolated from various environmental niches. *Environ. Microbiol.* **12:** 758-773.
9. Martino ME, Bayjanov JR, Caffrey BE, Wels M, Joncour P, Hughes S, *et al*. 2016. Nomadic lifestyle of *Lactobacillus plantarum* revealed by comparative genomics of 54 strains isolated from different habitats. *Environ. Microbiol.* **18:** 4974-4989.
10. Galloway-Peña J, Roh JH, Latorre M, Qin X, Murray BE. 2012. Genomic and SNP analyses demonstrate a distant separation of the hospital and community-associated clades of *Enterococcus faecium*. *PLoS One* **7:** e30187.
11. Kim EB, Marco ML. 2014. Nonclinical and clinical *Enterococcus faecium* strains, but not *Enterococcus faecalis* strains, have distinct structural and functional genomic features. *Appl. Environ. Microbiol.* **80:** 154-165.
12. Kopit LM, Kim EB, Siezen RJ, Harris LJ, Marco ML. 2014. Safety of the surrogate microorganism *Enterococcus faecium* NRRL B-2354 for use in thermal process validation. *Appl. Environ. Microbiol.* **80:** 1899-1909.
13. Kim EB, Jin GD, Lee JY, Choi YJ. 2016. Genomic features and niche-adaptation of *Enterococcus faecium* strains from Korean soybean-fermented foods. *PLoS One* **11:** e0153279.
14. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, *et al*. 2008. The RAST server: rapid annotations using subsystems technology. *BMC Genomics* **9:** 75.
15. Rizk G, Lavenier D. 2010. GASSST: global alignment short sequence search tool. *Bioinformatics* **26:** 2534-2540.
16. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32:** 1792-1797.
17. Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33:** 1870-1874.
18. Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406-425.
19. Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39:** 783-791.
20. Tamura K, Nei M, Kumar S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc. Natl. Acad. Sci. USA* **101:** 11030-11035.
21. Siliakus MF, van der Oost J, Kengen SW. 2017. Adaptations of archaeal and bacterial membranes to variations in temperature, pH and pressure. *Extremophiles* **21:** 651-670.
22. Aizenman E, Engelberg-Kulka H, Glaser G. 1996. An *Escherichia coli* chromosomal "addiction module" regulated by guanosine [corrected] 3′,5′-bispyrophosphate: a model for programmed bacterial cell death. *Proc. Natl. Acad. Sci. USA* **93:** 6059-6063.
23. Mittenhuber G. 1999. Occurence of MazEF-like antitoxin/toxin systems in bacteria. *J. Mol. Microbiol. Biotechnol.* **1:** 295-302.
24. Nyström T. 1998. To be or not to be: the ultimate decision of the growth-arrested bacterial cell. *FEMS Microbiol. Rev.* **21:** 283-290.
25. Gormley NA, Watson MA, Halford SE. 2001. Bacterial restriction–modification systems. *eLS* DOI: 10.1038/npg.els.0001037.
26. Kobayashi I. 2001. Behavior of restriction–modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res.* **29:** 3742-3756.
27. Plumed-Ferrer C, Koistinen KM, Tolonen TL, Lehesranta SJ, Karenlampi SO, Makimattila E, *et al*. 2008. Comparative study of sugar fermentation and protein expression patterns of two *Lactobacillus plantarum* strains grown in three different media. *Appl. Environ. Microbiol.* **74:** 5349-5358.